INSTITUTUL POLITEHNIC IAȘI
Facultatea de Electrotehnică
CATEDRA DE MATEMATICĂ

T. Bîrsan *I. Burdujan* *I. Vrabie*

METODE NUMERICE

— 1990 —

INSTITUTUL POLITEHNIC IASI
FACULTATEA DE ELECTROTEHNICA
CATEDRA DE MATEMATICA

T.BIRSAN I.BURDUJAN I.VRABIE

METODE NUMERICE

- 1989 -

PREFATA

Prezentul curs de "Metode numerice" este adresat cu precădere studenţilor din anul al II-lea ai facultăţii de Electrotehnică.

Constituindu-se ca o disciplină matematică, bazată în esenţă pe noţiunile şi rezultatele analizei şi algebrei pe care le reconsideră din punctul de vedere specific ei, metodele numerice au o pondere importantă în pregătirea viitorului inginer.

Destinat să introducă studentul în problematica specifică tehnicilor de aproximare, cursul urmează cu stricteţe programa cadru. Alegerea metodelor prezentate s-a făcut avîndu-se în vedere în primul rînd necesităţile obiectelor de specialitate, dar şi considerentele uzuale de ordin didactic.

Materialul a fost redactat după cum urmează:

Lect. dr. T. Bîrsan - capitolele 5, 6 şi 7,
Lect. dr. I. Burdujan - capitolele 2, 3 şi 4,
Lect. dr. I. Vrabie - capitolele 1 şi 8.

20 decembrie 1987

Autorii

Capitolul I

NOTIUNI INTRODUCTIVE

Scopul acestui capitol este de a pune la îndemîna cititorului noţiunile şi rezultatele auxiliare folosite frecvent pe parcursul întregii expuneri. Unele dintre aceste noţiuni şi rezultate au fost prezentate în cadrul cursurilor de: analiză matematică, algebră liniară, geometrie şi ecuaţii diferenţiale şi matematici speciale. Cu toate acestea, le-am inclus aici în ideea de a face lectura întregului material independentă de alte surse bibliografice.

§ 1. Inegalităţi

In acest paragraf prezentăm mai multe inegalităţi remarcabile deosebit de utile în cele ce urmează.

Lema 1.1. Pentru orice $p,q\in(1,+\infty)$ satisfăcînd $1/p+1/q=1$ şi orice $a,b\in[0,+\infty)$ are loc

$$ab\leq\frac{a^p}{p}+\frac{b^q}{q}. \tag{1.1}$$

Demonstraţie. Fie $f:[0,+\infty)\to[0,+\infty)$ definită prin $f(x)=\frac{x^p}{p}-x$ pentru orice $x\in[0,+\infty)$. Deoarece $f'(x)=x^{p-1}-1$ pentru orice $x\in[0,+\infty)$, se constată cu uşurinţă că $x=1$ este punctul de minim global pentru funcţia f. In consecinţă

$$f(1)\leq f(x)\ ,\quad (\forall)\ x\in[0,+\infty),$$

sau echivalent

$$\frac{1}{p}-1\leq\frac{x^p}{p}-x\ ,\ (\forall)x\in[0,+\infty)\ .$$

Ţinînd cont că $1-\frac{1}{p}=\frac{1}{q}$, ultima inegalitate se mai poate scrie

$$x\leq\frac{x^p}{p}+\frac{1}{q}\ ,\quad (\forall)x\in[0,+\infty)\ .$$

In particular, pentru $x = ab^{1-q}$ avem

$$ab^{1-q} \leq \frac{a^p \cdot b^{1-q}}{p} + \frac{1}{q}$$

Inmulţind ultima inegalitate în ambii membri cu b^q şi observînd că $p(1-q) + q = o$, obţinem (1.1).

<u>Consecinţa 1.1</u>. <u>Pentru orice</u> $p,q \in (1,+\infty)$ <u>satisfăcînd</u> $1/p + 1/q = 1$, <u>orice</u> $a,b \in [o,+\infty)$ <u>şi orice</u> $\varepsilon > o$ <u>are loc</u>

$$(1.2) \qquad ab \leq \varepsilon^p \frac{a^p}{p} + \frac{1}{\varepsilon^q} \cdot \frac{b^q}{q} .$$

<u>Demonstraţie</u>. Se aplică Lema 1.1 numerelor $a' = \varepsilon a$ şi $b' = \frac{1}{\varepsilon} b$.

<u>Observaţia 1.1</u>. Se constată cu uşurinţă că pentru $p=q=2$ inegalitatea (1.1) nu este altceva decît binecunoscuta inegalitate a mediilor, iar (1.2) inegalitatea lui Cauchy cu ε .

<u>Lema 1.2</u>. <u>(Inegalitatea lui Hölder pentru numere)</u>. <u>Fie</u> $n \in N^*$ şi $p,q \in [1,+\infty)$ <u>cu</u> $1/p + 1/q = 1$. <u>Atunci, pentru orice</u> $a_1, a_2, \ldots, a_n \in R$ <u>şi orice</u> $b_1, b_2, \ldots, b_n \in R$ <u>are loc</u>

$$(1.3) \qquad \sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^p\right)^{1/p} \left(\sum_{i=1}^{n} b_i^q\right)^{1/q}$$

<u>Demonstraţie</u>. Dacă $\sum_{i=1}^{n} a_i^p = o$ sau $\sum_{i=1}^{n} b_i^q = o$, atunci $a_1 = a_2 = \ldots = a_n = o$ sau $b_1 = b_2 = \ldots = b_n = o$ şi evident are loc (1.3). De aceea, putem presupune, fără a restrînge generalitatea, că $\sum_{i=1}^{n} a_i^p \neq o$ şi $\sum_{i=1}^{n} b_i^q \neq o$. Definim

$$x_i = \frac{a_i}{\left(\sum_{i=1}^{n} a_i^p\right)^{1/p}} \quad \text{şi} \quad y_i = \frac{b_i}{\left(\sum_{i=1}^{n} b_i^q\right)^{1/q}} , \ i=1,2,\ldots,n.$$

Din Lema 1.1 deducem

$$\sum_{i=1}^{n} x_i y_i \le \sum_{i=1}^{n} \frac{x_i^p}{p} + \sum_{i=1}^{n} \frac{y_i^q}{q} .$$

Inlocuim în această inegalitate x_i şi y_i în funcţie de a_i şi b_i, $i = 1,2,\dots,n$, obţinem

$$\Big(\sum_{i=1}^{n} a_i^p\Big)^{-1/p} \Big(\sum_{i=1}^{n} b_i^q\Big)^{-1/q} \sum_{i=1}^{n} a_i b_i \le \frac{1}{p} + \frac{1}{q} = 1 ,$$

inegalitate care este echivalentă cu (1.3).

<u>Lema 1.3.</u> (<u>Inegalitatea lui Hölder pentru funcţii</u>). <u>Fie</u> $p,q \in (1,+\infty)$ <u>cu</u> $1/p + 1/q = 1$. <u>Atunci, pentru orice funcţii integrabile Riemann</u> $f,g:[a,b] \to R$ <u>are loc</u>

$$\int_a^b f(t)g(t)dt \le \Big(\int_a^b |f(t)|^p dt\Big)^{1/p} \Big(\int_a^b |g(t)|^q dt\Big)^{1/q} \tag{1.4}$$

<u>Demonstraţie</u>. Fie $n \in N^*$ şi fie Δ_n diviziunea intervalului $[a,b]$, $\Delta_n : a = x_0 < x_1 < \dots < x_n = b$ cu $x_{i+1} - x_i = \frac{b-a}{n}$, $i=0,1,\dots,n-1$.

Fie $\xi_i \in [x_i, x_{i+1}]$, $i=0,1,\dots,n-1$ puncte intermediare fixate şi să aplicăm inegalitatea (1.3) sumei Riemann

$$\sum_{i=0}^{n} f(\xi_i)g(\xi_i)(x_{i+1} - x_i) = \frac{b-a}{n} \sum_{i=0}^{n} f(\xi_i)g(\xi_i).$$

Obţinem

$$\begin{aligned} \frac{b-a}{n} \sum_{i=0}^{n-1} f(\xi_i)g(\xi_i) &\le \frac{b-a}{n} \sum_{i=0}^{n-1} |f(\xi_i)|\,|g(\xi_i)| \le \\ &\le \frac{b-a}{n}\Big(\sum_{i=0}^{n-1} |f(\xi_i)|^p\Big)^{1/p} \Big(\sum_{i=0}^{n} |g(\xi_i)|^q\Big)^{1/q} = \\ &= \Big(\frac{b-a}{n}\Big)^{1/p} \Big(\frac{b-a}{n}\Big)^{1/q} \Big(\sum_{i=0}^{n-1} |f(\xi_i)|^p\Big)^{1/p} \Big(\sum_{i=0}^{n-1} |g(\xi_i)|^q\Big)^{1/q} = \\ &= \Big[\sum_{i=0}^{n-1} |f(\xi_i)|^p (x_{i+1} - x_i)\Big]^{1/p} \Big[\sum_{i=0}^{n-1} |g(\xi_i)|^q (x_{i+1} - x_i)\Big]^{1/q} \end{aligned}$$

In consecință

$$\sum_{i=0}^{n-1} f(\xi_i)g(\xi_i)(x_{i+1} - x_i) \leq$$

$$\leq \left[\sum_{i=0}^{n-1} |f(\xi_i)|^p(x_{i+1} - x_i)\right]^{1/p}\left[\sum_{i=0}^{n-1} |g(\xi_i)|^q(x_{i+1} - x_i)\right]^{1/q}.$$

Intrucît f și g sînt integrabile Riemann, $f.g, |f|^p$ și $|g|^q$ sînt de asemenea integrabile Riemann. Trecînd la limită în ultima inegalitate pentru $n \to +\infty$ obținem (1.4).

__Observația 1.2__. Pentru $p = q = 2$, (1.3) și (1.4) se reduc la inegalitatea lui Cauchy-Schwarz și Buniakowsky-Schwarz, respectiv.

__Lema 1.4__. __(Inegalitatea lui Minkowski pentru numere)__

__Fie__ $p \in [1,+\infty)$. __Atunci, pentru orice__ $a_1,a_2,\ldots,a_n \in R$ __și orice__ $b_1,b_2,\ldots,b_n \in R$ __are loc__

$$(1.5) \qquad \left(\sum_{i=1}^{n} |a_i + b_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{n} |a_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n}|b_i|^p\right)^{1/p}.$$

__Demonstrație__. Pentru $p = 1$ inegalitatea (1.5) este evidentă. De aceea, fie $p > 1$. Avem

$$\sum_{i=1}^{n} |a_i + b_i|^p = \sum_{i=1}^{n} |a_i + b_i||a_i + b_i|^{p-1} \leq$$

$$\sum_{i=1}^{n} |a_i||a_i + b_i|^{p-1} + \sum_{i=1}^{n} |b_i||a_i + b_i|^{p-1}.$$

Fie $q \in (1,+\infty)$ cu $\frac{1}{p} + \frac{1}{q} = 1$. Să remarcăm că un asemenea număr există și este unic determinat ($q = \frac{p}{p-1}$) întrucît $p > 1$. Din (1.3) rezultă atunci

$$\sum_{i=1}^{n}|a_i||a_i+b_i|^{p-1}+\sum_{i=1}^{n}|b_i||a_i+b_i|^{p-1}\leq$$

$$\left[\left(\sum_{i=1}^{n}|a_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|b_i|^p\right)^{1/p}\right]\left(\sum_{i=1}^{n}|a_i+b_i|^{pq-q}\right)^{1/q}$$

şi în consecinţă

$$\sum_{i=1}^{n}|a_i+b_i|^p\leq\left[\left(\sum_{i=1}^{n}|a_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|b_i|^p\right)^{1/p}\right]\left(\sum_{i=1}^{n}|a_i+b_i|^{pq-q}\right)^{1/q}.$$

Ţinînd cont că $pq-q=p$ şi că $1/q=1-1/p$ ultima inegalitate se scrie

$$\sum_{i=1}^{n}|a_i+b_i|^p\leq\left[\left(\sum_{i=1}^{n}|a_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|b_i|^p\right)^{1/p}\right]\left(\sum_{i=1}^{n}|a_i+b_i|^p\right)^{1-1/p},$$

inegalitate care, evident, implică (1.5).

<u>Lema 1.5</u>. (<u>Inegalitatea lui Minkowski pentru funcţii</u>). <u>Fie</u> $p\in[1,+\infty)$. <u>Atunci pentru orice funcţii integrabile Riemann</u> $f,g:[a,b]\to R$ <u>are loc</u>

$$(1.6)\qquad \left(\int_a^b|f(t)+g(t)|^p dt\right)^{1/p}\leq\left(\int_a^b|f(t)|^p dt\right)^{1/p}+\left(\int_a^b|g(t)|^p dt\right)^{1/p}$$

<u>Demonstraţie</u>. Fie $n\in N^*$ şi fie $\Delta_n: a=x_0<x_1<\ldots<x_n=b$ diviziunea intervalului $[a,b]$ cu $x_{i+1}-x_i=\frac{b-a}{n}$ pentru $i=0,1,\ldots$ $\ldots,n-1$. Fie $\xi_i\in[x_i,x_{i+1}]$ puncte intermediare fixate, şi să aplicăm inegalitatea (1.5) sumei Riemann

$$\left[\sum_{i=0}^{n}|f(\xi_i)+g(\xi_i)|^p(x_{i+1}-x_i)\right]^{1/p}=\left(\frac{b-a}{n}\right)^{1/p}\left(\sum_{i=0}^{n-1}|f(\xi_i)+g(\xi_i)|^p\right)^{1/p}$$

Obţinem

$$\left(\frac{b-a}{n}\right)^{1/p}\left(\sum_{i=0}^{n-1}|f(\xi_i)+g(\xi_i)|^p\right)^{1/p}\leq\left(\frac{b-a}{n}\right)^{1/p}\left[\left(\sum_{i=0}^{n-1}|f(\xi_i)|^p\right)^{1/p}+\right.$$

$$+ \Big(\sum_{i=0}^{n-1} |g(\xi_i)|^p\Big)^{1/p}\Big],$$

sau, echivalent

$$\Big[\sum_{i=0}^{n-1} |f(\xi_i)+g(\xi_i)|^p(x_{i+1}-x_i)\Big]^{1/p} \leqslant \Big[\sum_{i=0}^{n-1} |f(\xi_i)|^p(x_{i+1}-x_i)\Big]^{1/p} +$$

$$+\Big[\sum_{i=0}^{n-1} |g(\xi_i)|^p(x_{i+1}-x_i)\Big]^{1/p}.$$

Cum f şi g sînt integrabile Riemann, $|f+g|^p$, $|f|^p$ şi $|g|^p$ sînt de asemenea integrabile Riemann şi în consecinţă, trecînd la limită pentru $n \to +\infty$ în ultima inegalitate, obţinem (1.6).

§ 2. Integrala Lebesgue

Scopul acestui paragraf este de a defini un concept de integrală mai general decît cel de integrală Riemann în ideea de a lărgi clasa funcţiilor integrabile - pe de o parte - şi de a elimina unele neajunsuri ale integralei Riemann - pe de altă parte.

Definiţia 2.1. O submulţime E din R se numeşte neglijabilă, sau de măsură nulă, dacă pentru orice $\varepsilon > 0$ există o familie de intervale deschise $\{I_m;\ m \in N\}$ astfel încît

$$E \subseteq \bigcup_{n=0}^{\infty} I_n$$

şi

$$\sum_{n=0}^{\infty} \ell(I_n) < \varepsilon,$$

unde pentru orice $n \in N$, $\ell(I_n)$ este lungimea intervalului $I_n = (a_n, b_n)$, adică $\ell(I_n) = b_n - a_n$.

Observaţia 2.1. In definiţia de mai sus, admitem ca unele din intervalele I_n să fie vide, altfel spus, $I_n = (a_n, a_n)$.

In acest caz, $\ell(\phi) = \ell(I_n) = a_n - a_n = o$.

<u>Observaţia 2.2</u>. Orice mulţime finită $E = \{a_o, a_1, \dots, a_k\}$ este neglijabilă. In adevăr, dacă $\varepsilon > o$ este arbitrar, atunci familia de intervale $\{I_n;\ n \in N\}$ definită prin

$$I_n = (a_n - \frac{\varepsilon}{2(k+1)}, a_n + \frac{\varepsilon}{2(k+1)}), \quad \text{pentru} \quad n = o,1,2,\dots,k$$

şi

$$I_n = \phi, \text{ pentru } n > k$$

satisface condiţiile din Definiţia 2.1.

<u>Propoziţia 2.1</u>. <u>Fie</u> $\{E_n; n \in N\}$ <u>o familie de mulţimi neglijabile</u>.

<u>Atunci</u> $\bigcup_{n \in N} E_n$ <u>este de asemenea neglijabilă</u>

<u>Observaţia 2.3</u>. Condiţia ca familia de mulţimi neglijabile să fie numărabilă, adică indexată după N, este esenţială. Subliniem că, în general, dacă $\{E_\alpha; \alpha \in \Gamma\}$ este o familie de mulţimi neglijabile şi $\Gamma \neq N$, nu rezultă că $\bigcup_{\alpha \in \Gamma} E_\alpha$ este tot neglijabilă. De exemplu, dacă $\Gamma = [o,1]$ şi $E_\alpha = \{\alpha\}$ pentru orice $\alpha \in [o,1]$, atunci $\{E_\alpha; \alpha \in \Gamma\}$ conţine numai mulţimi neglijabile (fiind finite - vezi Observaţia 2.2) dar $\bigcup_{\alpha \in \Gamma} E_\alpha = [o,1]$ care, evident, nu este neglijabilă.

<u>Demonstraţia Propoziţiei 2.1.</u> Cum E_n este neglijabilă pentru $\varepsilon > o$, există o familie numărabilă de intervale $\{I_{n,k}; k \in N\}$ astfel încît

$$\bigcup_{k=o}^{\infty} I_{n,k} \supseteq E_n$$

şi

$$\sum_{k=o}^{\infty} \ell(I_{n,k}) < \varepsilon/2^{n+1}$$

Renumerotînd intervalele $\{I_{n,k};\ n \in N,\ k \in N\}$ în ordinea indicată de săgeţi - vezi tabelul de mai jos - se obţine familia $\{J_m;\ m \in N\} = \{I_{n,k};\ n \in N,\ k \in N\}$.

$$\begin{array}{cccccccc}
I_{1,1} \rightarrow I_{1,2} & I_{1,3} & \rightarrow - - - & I_{1,k} & I_{1,k+1} & - - - \\
\swarrow \quad \nearrow \quad \swarrow & & & & & \\
I_{2,1} \quad I_{2,2} & I_{2,3} & - - - - & I_{2,k}, & I_{2,k+1} & - - - \\
\downarrow \quad \nearrow \quad \swarrow & & & & & \\
I_{3,1} \quad I_{3,2} & I_{3,3} & - - - - & I_{3,k}, & I_{3,k+1} & - - - \\
\vdots \quad \swarrow \quad \nearrow & & & & & \\
I_{n,1} \quad I_{n,2} & I_{n,3} & - - - - & I_{n,k} & I_{n,k+1} & - - - \\
\vdots & & & & &
\end{array}$$

cu $J_1 = I_{1,1}$, $J_2 = I_{1,2}$, $J_3 = I_{2,1}$, $J_3 = I_{3,1}$ ş.a.m.d.

Evident $\quad \bigcup_{n=o}^{\infty} E_n \subseteq \bigcup_{m=o}^{\infty} J_m = \bigcup_{n=o}^{\infty} \bigcup_{k=o}^{\infty} I_{n,k}$ şi

$$\sum_{m=o}^{\infty} \ell(I_m) = \sum_{n=o}^{\infty} \sum_{k=o}^{\infty} \ell(I_{m,k}) < \varepsilon .$$

Observaţia 2.4. Procedeul de renumerotare (reindexare) descris în demonstraţia Propoziţiei 2.1 poartă numele de procedeul diagonal al lui Cantor. Cu ajutorul lui se poate arăta că mulţimea Q poate fi scrisă sub forma $Q = \{x_n;\ n \in N\}$.

Intr-adevăr, avem

$$\begin{array}{cccccccccccc}
o \rightarrow 1 & -1 \rightarrow \frac{1}{2} & -\frac{1}{2} & \frac{1}{3} & -\frac{1}{3} & - - - & \frac{1}{n} & -\frac{1}{n} & - - - \\
2 \quad -2 & \frac{2}{3} & -\frac{2}{3} & \frac{2}{5} & -\frac{2}{5} & \frac{2}{7} & -\frac{2}{7} & - - - - - - - - \\
3 \quad -3 & \frac{3}{2} & -\frac{3}{2} & \frac{3}{4} & -\frac{3}{4} & \frac{3}{5} & -\frac{3}{5} & - - - - - - - - - , \\
\vdots & & & & & & & &
\end{array}$$

ceea ce demonstrează afirmaţia de mai sus.

Din Propoziţia 2.1 rezultă atunci că mulţimea $Q = \bigcup_{n \in N} \{x_n\}$ este neglijabilă.

Definiţia 2.2. Fie $E \subset R$. Numim funcţie caracteristică a mulţimii E, funcţia $\chi_E : R \rightarrow R$ definită prin

$$\chi_E(x) = \begin{cases} 1 & \text{dacă } x \in E \\ o & \text{dacă } x \in R \setminus E. \end{cases}$$

In tot ceea ce urmează vom întrebuința denumirea de interval pentru a desemna intervale de forma $[a,b)$ cu $a,b \in R$, $a \le b$.

Definiția 2.3. Se numește funcție etajată orice combinație liniară de caracteristici de intervale.

In consecință, o funcție etajată $f:R \to R$ are forma

$$(2.1) \qquad f(x) = \sum_{i=1}^{n} f_i \chi_{E_i}(x) \qquad (\forall)\ x \in R ,$$

unde E_i sînt intervale, $i = 1,2,\ldots,n$.

Evident scrierea (2.1) nu este unică deoarece orice interval poate fi scris ca o reuniune de mai multe intervale mutual disjuncte.

Propoziția 2.2. Orice funcție etajată poate fi scrisă sub forma (2.1) cu $E_i \cap E_j = \emptyset$ pentru $i \neq j$.

Demonstrație. Fie $f(x) = f_1 \chi_{E_1}(x) + f_2 \chi_{E_2}(x)$ $(\forall) x \in R$.

Notînd cu

$$F_1 = E_1 \cap (R \setminus E_2),\ F_2 = E_1 \cap E_2 \text{ și } F_3 = E_2 \cap (R \setminus E_1)$$

$$\bar{f}_1 = f_1,\ \bar{f}_2 = f_1 + f_2,\ \bar{f}_3 = f_2 ,$$

avem

$$f(x) = \bar{f}_1 \chi_{F_1}(x) + \bar{f}_2 \chi_{F_2}(x) + \bar{f}_3 \chi_{F_3}(x) \quad (\forall)\ x \in R,$$

unde F_1, F_2, F_3 sînt intervale mutual disjuncte.

Demonstrația cazului general se reduce la repetarea raționamentului de mai sus de un număr finit de ori, din aproape în aproape.

Propoziția 2.3. a) Orice combinație liniară de funcții etajate este o funcție etajată. Deci mulțimea $\mathcal{E}$ a funcțiilor etajate este un spațiu vectorial (liniar) peste R.

b) Dacă $f \in \mathcal{E}$ atunci $|f| \in \mathcal{E}$.

c) Dacă $f \in \mathcal{E}$ atunci $f_+ = \frac{f+|f|}{2}$ și $f_- = \frac{f-|f|}{2}$ aparțin lui $\mathcal{E}$.

d) Dacă $f, g \in \mathcal{E}$ atunci $f \cdot g \in \mathcal{E}$.

e) Dacă $f_1, f_2, \ldots, f_n \in \mathcal{E}$ atunci funcţiile $M, m : R \to R$ definite prin

$$M(x) = \max \{ f_i(x);\ i = \overline{1,n} \} \quad (\forall)\ x \in R$$

$$m(x) = \min \{ f_i(x);\ i = \overline{1,n} \} \quad (\forall)\ x \in R$$

aparţin lui $\mathcal{E}$.

Demonstraţia fiind foarte simplă o lăsăm în seama cititorului.

Pentru orice interval $E = [a,b)$ notăm cu $m(E) = b-a$ măsura sau lungimea lui. Avem $m(\emptyset) = m([a,a)) = 0$.

Propoziţia 2.4. Dacă intervalul $E = [a,b)$ este reuniunea unei familii de intervale $\{E_n;\ n \in N$ cu $E_i \cap E_j = \emptyset$ pentru orice $i \neq j$, atunci

$$m(E) = \sum_{n=0}^{\infty} m(E_n).$$

Demonstraţie. Schimbînd eventual numerotarea, putem presupune fără a restrînge generalitatea că

$$E_1 = [a, b_1),\ E_2 = [b_1, b_2), \ldots, E_n = [b_{n-1}, b_n), \ldots$$

Evident

$$\sum_{n=1}^{k} m(E_n) = b_k - a \leq b - a = m(E)$$

şi atunci seria $\sum_{n=1}^{\infty} m(E_n)$ este convergentă şi

$$\sum_{n=1}^{\infty} m(E_n) \leq m(E)$$

Observînd că $\sup \{ b_k;\ k \in N \} = b$, obţinem

$$\sum_{n=1}^{\infty} m(E_n) = \lim_{k \to \infty} (b_k - a) = b - a = m(E)$$

Propoziţia 2.5. Fie $f \in \mathcal{E}$. Dacă

$$f(x) = \sum_{i=1}^{n} f_i \chi_{E_i}(x) \quad (\forall)\ x \in R$$

şi

$$f(x) = \sum_{j=1}^{m} \tilde{f}_j \chi_{F_j}(x), \quad (\forall)\ x \in R$$

sînt două reprezentări ale lui f cu $E_i \cap E_k = \emptyset$ pentru $i \neq k$ şi $F_j \cap F_h = \emptyset$ pentru $j \neq h$ atunci

$$\sum_{i=1}^{n} f_i\, m(E_i) = \sum_{j=1}^{m} \tilde{f}_j\, m(F_j)$$

Demonstraţie. Se aplică Propoziţia 2.4.

Definiţia 2.4. Fie $f \in \mathcal{E}$ şi fie $f(x) = \sum_{i=1}^{n} f_i \chi_{E_i}(x), (\forall) x \in R$ o reprezentare a lui f cu $E_i \cap E_k = \emptyset$ pentru $i \neq k$. Vezi Propoziţia 2.2. Numim integrala funcţiei f numărul

$$I(f) = \sum_{i=1}^{n} f_i\, m(E_i).$$

Din Propoziţia 2.5 rezultă că numărul $I(f)$ este independent de reprezentarea lui f ca o combinaţie liniară de funcţii caracteristice de intervale mutual disjuncte. In plus, cum orice f este nulă în afara unui interval $[a,b]$, se constată cu uşurinţă că

(2.2) $\quad I(f) = \int_a^b f(t)dt$ - integrala Riemann a funcţiei f.

Propoziţia 2.6. Au loc următoarele proprietăţi

a) Pentru orice $f, g \in \mathcal{E}$ şi orice $\alpha, \beta \in R$ avem

$$I(\alpha f + \beta g) = \alpha I(f) + \beta I(g)$$

b) $I(o) = o$ şi $I(f) > o$ pentru orice $f \geq o$, $f \neq o$

c) Dacă $f, g \in \mathcal{E}$ şi $f(x) \leq g(x)$, $(\forall)$ $x \in R$ atunci $I(f) \leq I(g)$.

d) Pentru orice $f \in \mathcal{E}$ avem $|I(f)| \leq I(|f|)$.

Lăsăm demonstraţia acestor proprietăţi simple, dar fundamentale, pe seama cititorului. A se vedea (2.2).

Definiţia 2.5. Un şir de funcţii etajate (f_m) se numeşte şir Cauchy dacă $(\forall)$ $\varepsilon > o$ $(\exists)$ $n(\varepsilon) \in N$ astfel încît pentru orice $n \in N$, $n \geq n(\varepsilon)$ şi orice $p \in N$ avem

$$I(|f_n - f_{n+p}|) \leq \varepsilon.$$

Evident, $(f_n) \subset \mathcal{E}$ este, şir Cauchy dacă şi numai dacă

$$\lim_{\substack{n \to +\infty \\ m \to +\infty}} I(|f_n - f_m|) = o,$$

unde n şi m tind independent la $+\infty$.

Teorema 2.1. Din orice şir Cauchy de funcţii etajate se poate extrage un subşir cu proprietatea că pentru orice $\varepsilon > o$ există o familie $\{I_n^{\varepsilon}\ ;\ n \in N\}$ de intervale cu $\sum_{n=o}^{\infty} m(I_n^{\varepsilon}) \leq \varepsilon$, astfel încît subşirul converge uniform pe $R \setminus \bigcup_{n=o}^{\infty} I_n^{\varepsilon}$ la o funcţie $f: R \to R$.

Demonstraţie. Să observăm că dacă $f \in \mathcal{E}$ şi $\varepsilon \geq o$ atunci $\{x \in R;\ |f(x)| \geq \varepsilon\}$ poate fi reprezentată ca o reuniune numărabilă de intervale mutual disjuncte, adică,

$$\{x \in R;\ |f(x)| \geq \varepsilon\} = \bigcup_{n=o}^{\infty} I_n \text{ cu } I_n \cap I_m = \emptyset \text{ pentru } n \neq m.$$

Notînd cu $E_{\varepsilon} = \{x \in R;\ |f(x)| < \varepsilon\}$ şi cu $F_{\varepsilon} = \{x \in R;\ |f(x)| \geq \varepsilon\}$, să observăm că

$$I(|f|) = I(|f|\chi_{E_{\varepsilon}}) + I(|f|\chi_{F_{\varepsilon}}) \geq I(|f|\chi_{F_{\varepsilon}}) \geq \varepsilon\, m(F_{\varepsilon})$$

In consecințǎ avem

$$\varepsilon \cdot m(\{x \in R;\ |f(x)| \geq \varepsilon\}) \leq I(|f|) \tag{2.3}$$

Fie (f_n) un șir Cauchy de funcții etajate. Din Definiția 2.5 rezultă că pentru orice $k \in N$ există $n_k = n(2^{-2k})$ astfel încît

$$I(|f_n - f_m|) \leq 2^{-2k} \text{ pentru orice } n \geq n_k \text{ și } m \geq n_k \tag{2.4}$$

Evident, putem presupune, fără a restrînge generalitatea, că $n_k \leq n_{k+1}$ pentru orice $k \in N$.

Notînd cu

$$E_k = \{x \in R;\ |f_{n_k}(x) - f_{n_{k+1}}(x)| \geq 2^{-k}\}$$

și cu

$$F_k = \{x \in R;\ |f_{n_k}(x) - f_{n_{k+1}}(x)| < 2^{-k}\} = R \setminus E_k ,$$

avem (vezi (2.3) și (2.4))

$$m(\{x \in R;\ |f_{n_k}(x) - f_{n_{k+1}}(x)| \geq 2^{-k}\}) \leq 2^{-k} \tag{2.5}$$

și

$$|f_{n_k}(x) - f_{n_{k+1}}(x)| < 2^{-k} \quad (\forall) x \in F_k = R \setminus E_k . \tag{2.6}$$

Fie $I_k = \bigcup_{n=k}^{\infty} E_n$ pentru orice $k \in N$. Intrucît E_n este o reuniune numărabilă de intervale mutual disjuncte, I_k este tot o reuniune numărabilă de intervale. In plus

$$m(I_k) = m\left(\bigcup_{n=k}^{\infty} E_n\right) \leq \sum_{n=k}^{\infty} m(E_n) \leq \sum_{n=k}^{\infty} 2^{-n} = 2^{-k+1} \tag{2.7}$$

pentru orice $k \in N$.

Fie $\varepsilon > 0$ și fie $k_0 \in N$ astfel încît $2^{-k_0+1} \leq \varepsilon$. Din (2.6) deducem

$$(2.8)\quad |f_{n_k}(x)-f_{n_{k'}}(x)|\leq |f_{n_k}(x)-f_{n_{k+1}}(x)|+|f_{n_{k+1}}(x)-f_{n_{k+2}}(x)|+\ldots$$

$$+\ldots+|f_{n_{k'-1}}(x)-f_{n_{k'}}(x)|\leq 2^{-k}+2^{-k-1}+\ldots+2^{-k'}<2^{-k+1}$$

pentru orice k şi k', $k\geq k_o$, $k'\geq k_o$ şi orice $x\in R\setminus I_{k_o}$.

Dar inegalitatea (2.8) ne arată că (f_{m_k}) este un şir uniform Cauchy pe $R\setminus I_{k_o}$ şi ca atare el este convergent pe $R\setminus I_{k_o}$.

In plus $m(I_{k_o})\leq e^{-k_o+1}\leq \varepsilon$. Vezi (2.7).

Fie acum $x\in R\setminus\bigcap_{p=o}^{\infty} I_p=\bigcup_{p=o}^{\infty}(R-I_p)$. Există atunci $p_o\in N$ astfel încît $x\in R\setminus I_{p_o}$. Raţionînd ca mai sus deducem că

$$|f_{n_k}(x)-f_{n_{k'}}(x)|\leq 2^{-k+1}$$

pentru orice k, $k'\in N$ $k\geq p_o$ şi $k'\geq p_o$.

Deci (f_{n_k}) este şir <u>punctual</u> Cauchy pe $R\setminus\bigcap_{p=o}^{\infty} I_p$. Ca atare există o funcţie $g\colon R\setminus\bigcap_{p=o}^{\infty} I_p\to R$ astfel încît

$$\lim_{k\to+\infty} f_{n_k}(x)=g(x)$$

pentru orice $x\in R\setminus\bigcap_{p=o}^{\infty} I_p$. Să definim acum $f\colon R\to R$ prin

$$f(x)=\begin{cases} g(x) & \text{dacă } x\in R\setminus\bigcap_{p=o}^{\infty} I_p\\ o & \text{dacă } x\in\bigcap_{p=o}^{\infty} I_p\,. \end{cases}$$

Evident f satisface condiţiile teoremei. In plus

$$m\Big(\bigcap_{p=o}^{\infty} I_p\Big)\leq m(I_p)\leq 2^{-p+1}$$

pentru orice $p\in N$ şi deci $m\Big(\bigcap_{p=o}^{\infty} I_p\Big)=o$.

Dacă o anumită proprietate are loc pe complementara unei mulţimi neglijabile, spunem că acea proprietate are loc <u>aproape peste tot</u> - pe scurt a.p.t. De exemplu, $f\colon R\to R$ este a.p.t. continuă dacă există $E\subset R$ neglijabilă astfel încît f să fie

continuă pe $R \setminus E$.

Să observăm de asemenea, că subşirul (f_{n_k}) construit în demonstraţia Teoremei 2.1 este a.p.t. convergent la funcţia $f: R \to R$ definită ca mai sus.

Propoziţia 2.7. Dacă (f_n) este un şir Cauchy de funcţii etajate convergent a.p.t. la o, atunci $\lim_{n\to+\infty} I(f_n) = o$.

Demonstraţie. Din (a) şi (d) Propoziţia 2.6 avem

$|I(f_n) - I(f_{n+1})| \leq I(|f_n - f_{n+p}|)$ pentru orice $n, p \in N$ şi ca atare $(I(f_n))$ este un şir Cauchy de numere reale. Deci $(I(f_n))$ este convergent.

Pentru a completa demonstraţia este atunci suficient să arătăm că există un subşir al lui $(I(f_n))$ convergent la o.

In acest scop fie (f_{n_k}) subşirul dat de Teorema 2.1, pe care îl vom nota, pentru simplitate, tot cu (f_n).

Dacă există $n_o \in N$ astfel încît $f_n = o$ pentru orice $n \geq n_o$ demonstraţia este încheiată. Vezi (b), Propoziţia 2.6. Să presupunem atunci că există numere $n_o \in N$ - oricît de mari - pentru care $f_{n_o} \neq o$. Fie $\varepsilon > o$ şi fie $n_o \in N$ un număr cu proprietatea că $f_{n_o} \neq o$ şi

(2.9) $$I(|f_n - f_{n'}|) \leq \varepsilon/4 \text{ pentru orice } n, n' \geq n_o.$$

Fie $E = \{x \in R;\ f_{n_o}(x) \neq o\}$. Evident E se poate scrie ca o reuniune finită de intervale şi $m(E) > o$. Să notăm cu $S = \sup\{|f_{n_o}(x)|;\ x \in R\}$ şi să observăm că, din Teorema 2.1, există o reuniune numărabilă de intervale U cu $m(U) \leq \varepsilon/4.S$, astfel încît (f_n) converge uniform la o pe $R \setminus U$. In consecinţă, pentru $n \geq n_o$ - suficient de mare - avem

$$|f_n(x)| \leq \varepsilon/4\, m(E) \quad (\forall)\ x \in R - U.$$

Fie $E_1 = \{x \in R;\ |f_n(x)| > \varepsilon/4\ m(E)\}$. Evident $E_1 \subset U$ şi ca atare

(2.10) $$m(E_1) \leq \varepsilon/4\ S.$$

Să observăm că $R = E_1 \cup E_2 \cup E_3$ unde $E_2 = E \cap (R - E_1)$ şi $E_3 = R - E$ şi că E_1, E_2, E_3 sînt mutual disjuncte.

Rezultă atunci că f_n se poate descompune sub forma

$$f_n = (f_n - f_{n_o})\chi_{E_1} + f_{n_o}\chi_{E_1} + f_n\chi_{E_2} + f_n\chi_{E_3}$$

Intrucît $f_{n_o} = o$ pe E_3, $f_n = f_n - f_{n_o}$ pe E_3 şi atunci

$$f_n = (f_n - f_{n_o})\chi_{E_1} + f_{n_o}\chi_{E_1} + f_n\chi_{E_2} + (f_n - f_{n_o})\chi_{E_3}.$$

In consecinţă,

$$I(f_n) = I((f_n - f_{n_o})\chi_{E_1}) + I(f_{n_o}\chi_{E_1}) + I(f_n\chi_{E_2}) + I((f_n - f_{n_o})\chi_{E_3}),$$

şi din Propoziţia 2.6 deducem

$$|I(f_n)| \leq I(|f_n - f_{n_o}|) + |I(f_{n_o}\chi_{E_1})| + |I(f_n\chi_{E_2})| + I(|f_n - f_{n_o}|).$$

Intrucît din (2.9) avem $I(|f_n - f_{n_o}|) \leq \varepsilon/4$ pentru orice $n \geq n_o$ iar

$$|I(f_{n_o}\chi_{E_1})| \leq I(|f_{n_o}\chi_{E_1}|) \leq S.m(E_1) \leq S.\varepsilon/4\ S = \varepsilon/4 ,$$

(vezi (2.10)) şi

$$|I(f_n\chi_{E_2})| \leq I(|f_n\chi_{E_2}|) \leq [\varepsilon/4\ m(E)].m(E) = \varepsilon/4 ,$$

rezultă

$$|I(f_n)| \leq \varepsilon \text{ pentru orice } n \geq n_o.$$

Din $\lim\limits_{n\to\infty} I(f_n) = o$ şi demonstraţia este încheiată.

Propoziţia 2.8. Dacă şirul Cauchy de funcţii etajate (f_n) este convergent a.p.t. către o funcţie f, atunci şirul $(I(f_n))$ converge

la un număr real $I(f)$ care depinde numai de f nu şi de (f_n).

Demonstraţie. Din (a) şi (d) Propoziţia 2.6 avem

$$|I(f_n)-I(f_{n+p})|\leq I(|f_n-f_{n+p}|) \text{ pentru orice } n,p\in N$$

şi ca atare $(I(f_n))$ este un şir Cauchy de numere reale.

Există atunci

$$\lim_{n\to+\infty} I(f_n) = I(f).$$

Dacă $(\tilde{f}_n)$ este un alt şir Cauchy de funcţii etajate convergent a.p.t. la f, şirul $(f_n-\tilde{f}_n)$ este a.p.t. convergent la o şi este Cuachy. Din Propoziţia 2.7 rezultă atunci că

$$\lim_{n\to+\infty} I(f_n-\tilde{f}_n) = o\ ,$$

sau, echivalent

$$\lim_{n\to+\infty} I(f_n) = \lim_{n\to+\infty} I(\tilde{f}_n) = I(f)$$

Definiţia 2.6. O funcţie reală $f: R\to R$ se numeşte integrabilă Lebesgue (pe R) sau L-integrabilă dacă este limita a.p.t. a unui şir Cauchy de funcţii etajate. Numărul $I(f)$ dat de Propoziţia 2.8 se numeşte integrala Lebesgue a funcţiei f şi se notează cu

$$L\int_R f(t)dt \quad \text{sau} \quad \int_R f(t)dt \quad \text{sau} \quad \int_{-\infty}^{+\infty} f(t)dt.$$

Să observăm că două funcţii integrabile Lebesgue egale a.p.t. au aceeaşi integrală. De aceea, convenim să numim echivalente orice două funcţii integrabile Lebesgue egale a.p.t. Această relaţie este într-adevăr o relaţie de echivalenţă, iar mulţimea claselor de echivalenţă va fi notată cu $L^1(R)$. Prin abuz de limbaj vom scrie uneori că o funcţie $f\in L^1(R)$ înţelegînd prin aceasta că f aparţine unei clase de echivalenţă din $L^1(R)$.

Propoziţia 2.9. Dacă f este integrabilă Lebesgue şi (f_n) este un şir Cauchy de funcţii etajate care converge a.p.t. la f, atunci

$$\lim_{n\to+\infty}\int_R |f_n(t) - f(t)|dt = 0$$

Demonstraţie. Fie $\varepsilon > 0$. Există atunci $n(\varepsilon)\in N$ astfel încît

$$\int_R |f_n(t) - f_{n+p}(t)|dt \leq \varepsilon$$

pentru orice $n \geq n(\varepsilon)$ şi orice $p\in N$. Trecînd la limită pentru $p \to +\infty$ obţinem (vezi Propoziţia 2.8)

$$\int_R |f_n(t) - f(t)|dt \leq \varepsilon \text{ pentru orice } n \geq n(\varepsilon),$$

inegalitate care încheie demonstraţia.

Să observăm că dacă f este integrabilă Lebesgue atunci $|f|$ are aceeaşi proprietate. Vezi Propoziţia 2.3 (b) şi Definiţia 2.6. Putem atunci extinde noţiunea de şir Cauchy, de la mulţimea funcţiilor etajate, la $L^1(R)$.

Definiţia 2.7. Un şir (f_n) de funcţii integrabile Lebesgue se numeşte şir Cauchy dacă pentru orice $\varepsilon > 0$ există $n(\varepsilon)\in N$ astfel încît pentru orice $n \geq n(\varepsilon)$ şi orice $p\in N$ să avem

$$\int_R |f_n(t) - f_{n+p}(t)|dt \leq \varepsilon.$$

Teorema 2.2. Pentru orice şir Cauchy de funcţii integrabile Lebesgue (f_n) există o funcţie integrabilă Lebesgue f astfel încît

$$\lim_{n\to+\infty}\int_R |f_n(t) - f(t)|dt = 0$$

şi un subşir al lui (f_n) convergent a.p.t. la f.

Dacă g este o altă funcţie integrabilă Lebesgue cu proprietăţile de mai sus atunci f = g a.p.t.

Demonstraţie. Din Definiţia 2.6 şi Propoziţia 2.9 rezultă că pentru orice $n\in N$ există o funcţie etajată g_n astfel încît

$$\int_R |f_n(t) - g_n(t)|dt \leq 1/n.$$

Să observăm că

$$\int_R |g_n(t) - g_{n+p}(t)|dt \leq \int_R |g_n(t) - f_n(t)|dt + \int_R |f_n(t) - f_{n+p}(t)|dt +$$

$$+ \int_R |f_{n+p}(t) - g_{n+p}(t)|dt \leq 1/n + \int_R |f_n(t) - f_{n+p}(t)|dt + 1/(n+p)$$

pentru orice $n, p \in N$. Cum (f_n) este şir Cauchy urmează că şi (g_n) este şir Cauchy. Din Teorema 2.1 deducem că există o funcţie f şi un subşir al şirului (g_n) notat cu (g_{n_k}) astfel încît

$$\lim_{k \to +\infty} g_{n_k}(t) = f(t) \quad \text{a.p.t.} \quad t \in R.$$

In plus, conform Propoziţiei 2.8, f este integrabilă.

Pentru orice $n \in N$ fie n' indicele n_k cel mai apropiat de n. Evident n' tinde la infinit ori de cîte ori n tinde la infinit.

Să observăm că

$$\int_R |f_n(t) - f(t)|\,dt \leq \int_R |f_n(t) - g_n(t)|\,dt + \int_R |g_n(t) - g_{n'}(t)|dt +$$

$$+ \int_R |g_{n'}(t) - f(t)|dt \quad (\forall)\ n \in N,$$

şi în consecinţă

$$\lim_{n \to +\infty} \int_R |f_n(t) - f(t)|\,dt = 0.$$

Pentru a încheia demonstraţia, să remarcăm că din Teorema 2.1 rezultă că şirul (g_n) poate fi ales astfel încît

$$|g_n(t) - f_n(t)| \leq 2^{-n}$$

pentru orice $t \in R \setminus U_n$ unde U_n este o reuniune numărabilă de intervale cu suma măsurilor mai mici decît 2^{-n}. In acest caz

$$\lim_{n \to +\infty} (f_n - g_n) = 0$$

uniform pe complementara fiecărei mulţimi $V_k = \bigcup_{p=k}^{\infty} U_p$.

Evident V_k este reuniune numărabilă de intervale cu suma

lungimilor mai mici decît $2^{-k}+2^{-k-1}+\dots=2^{-k+1}$. Pe de altă parte, tot din Teorema 2.1, există un subşir (g_{n_p}) al lui (g_n) astfel încît, pentru orice $k\in N$

$$\lim_{p\to+\infty} g_{n_p}=f \quad \text{uniform pe } R\setminus W_k\,,$$

unde W_k este o reuniune numărabilă de intervale cu suma lungimilor mai mică decît $1/k$. In consecinţă, pentru orice $k\in N$ avem

$$\lim_{p\to+\infty}(f_{n_p}-f)=\lim_{p\to+\infty}(f_{n_p}-g_{n_p})+\lim_{p\to+\infty}(g_{n_p}-f)=0$$

pentru orice $t\in R\setminus V_k\cup W_k$. Intrucît $\bigcap_{k\in N}(V_k\cup W_k)$ este neglijabilă, rezultă că

$$\lim_{p\to+\infty} f_{n_p}=f \quad \text{a.p.t. } t\in R\,.$$

In sfîrşit, dacă (f_{n_q}) este un alt subşir cu

$$\lim_{q\to+\infty} f_{n_q}=g \quad \text{a.p.t. } t\in R\,,$$

din modul în care a fost definită funcţia f (vezi Teorema 2.1) urmează $f=g$ a.p.t. $t\in R$, ceea ce încheie demonstraţia.

<u>Consecinţa 2.1</u>. <u>Dacă</u> $f: R\to R_+$ <u>este o funcţie integrabilă Lebesgue şi</u>

$$\int_R f(t)dt=0$$

<u>atunci</u> $f(t)=0$ <u>a.p.t.</u> $t\in R$.

<u>Demonstraţie</u>. Se aplică Teorema 2.2 cu $f_n=f$ pentru orice $n\in N$. Atunci, cel puţin pe un subşir, avem

$$\lim_{n\to+\infty} f_n(t)=0 \quad \text{a.p.t. } t\in R\,,$$

ceea ce implică $f(t)=0$ a.p.t. $t\in R$.

<u>Teorema 2.3</u>. <u>(Beppo Levi) Dacă şirul de funcţii integrabile Lebesgue</u> (f_n) <u>este crescător a.p.t. şi şirul</u> $\left(\int_R f_n(t)dt\right)$ <u>este majorat atunci</u> (f_n) <u>converge a.p.t. către o funcţie integrabilă</u>

Lebesgue f și

$$\lim_{n\to+\infty}\int_R |f_n(t)-f(t)|dt = 0$$

In particular

$$\lim_{n\to+\infty}\int_R f_n(t)dt = \int_R f(t)dt.$$

Demonstrație. Intrucît (f_n) este crescător a.p.t. avem

$$|f_{n+p}(t)-f_n(t)| = f_{n+p}(t)-f_n(t) \quad \text{a.p.t.} \quad t\in R$$

și pentru orice $n,p\in N$. Cum $(\int_R f_n(t)dt)$ este crescător și majorat el este convergent și atunci

$$\int_R |f_{n+p}(t)-f_n(t)|dt = \int_R [f_{n+p}(t)-f_n(t)]dt = \int_R f_{n+p}(t)dt - \int_R f_n(t)dt$$

pentru orice $n,p\in N$. In consecință (f_n) este un șir Cauchy și ca atare, din Teorema 2.2 rezultă că există un subșir (f_{n_k}) convergent a.p.t. la o funcție integrabilă Lebesgue f. Din faptul că (f_n) este crescător a.p.t. deducem că (f_n) este convergent a.p.t. la f, iar din Teorema 2.2 avem

$$\lim_{n\to+\infty}\int_R |f_n(t)-f(t)|dt = 0$$

ceea ce implică

$$\lim_{n\to+\infty}\int_R f_n(t)dt = \int_R f(t)dt .$$

Teorema 2.4 (de convergență dominată a lui Lebesgue). Dacă șirul de funcții integrabile Lebesgue (f_n) este convergent a.p.t. către funcția f și există o funcție integrabilă Lebesgue g astfel încît $|f_n(t)|\le g(t)$ a.p.t. $t\in R$ pentru orice $n\in N$, atunci f este integrabilă Lebesgue și

$$\lim_{n\to+\infty}\int_R |f_n(t)-f(t)|dt = 0 .$$

In particular

$$\lim_{n\to+\infty}\int_R f_n(t)dt = \int_R f(t)dt .$$

Demonstrație. Conform Teoremei 2.2 este suficient să arătăm că (f_n) este șir Cauchy. In acest scop, să definim funcțiile

$$\varphi_{k,p}(t) = \sup\{|f_i(t) - f_j(t)|;\ \ k \le i,\ j \le p\}.$$

Pentru k fixat, șirul $(\varphi_{k,p})_{p\ge k}$ este crescător a.p.t. și în plus

$$\int_R \varphi_{k,p}(t)dt \le 2\int_R g(t)dt .$$

Să definim

$$\psi_k(t) = \sup\{\varphi_{k,p}(t);\ \ k \le p\}.$$

Din Teorema 2.3 rezultă că ψ_k este integrabilă pentru orice $k\in N$. In plus, (ψ_k) este descrescător a.p.t. și $\int_R \psi_k(t)dt \ge 0$ pentru orice $k\in N$. Intrucît $\lim_{n\to+\infty} f_n(t) = f(t)$ a.p.t. $t\in R$ urmează că $\lim_{k\to+\infty}\psi_k(t) = 0$ a.p.t. Din Teorema 2.3 aplicată șirului $(-\psi_k)$ rezultă

$$\lim_{k\to+\infty}\int_R \psi_k(t)dt = 0 .$$

Observînd că

$$\int_R |f_{n+p}(t) - f_n(t)|dt \le \int_R \psi_k(t)dt$$

pentru orice $n, p, k\in N$ cu $k \le n+p$, conchidem că (f_n) este șir Cauchy ceea ce încheie demonstrația.

Definiția 2.8. O funcție f definită pe o submulțime $E\subset R$ se numește integrabilă Lebesgue pe E dacă funcția $\hat{f} : R\to R$ definită prin

$$\hat{f}(t) = \begin{cases} f(t) & \text{dacă } t\in E \\ 0 & \text{dacă } t\in R\setminus E \end{cases}$$

este integrabilă Lebesgue pe R. Numărul

$$\int_E f(t)dt := \int_R \hat{f}(t)dt$$

poartă numele de <u>integrala Lebesgue</u> a funcţiei f pe mulţimea E

Dacă $E = [a,b]$ sau $E = R$ şi $p \geq 1$ vom nota cu $L^p[a,b]$, respectiv $L^p(R)$, mulţimea claselor de echivalenţă - în raport cu egalitatea a.p.t. - de funcţii $f:[a,b] \rightarrow R$, respectiv $f:R \rightarrow R$, integrabile Lebesgue cu proprietatea că şi $|f|^p$ este integrabilă Lebesgue.

In paragraful 4 vom pune în evidenţă mai multe proprietăţi remarcabile mulţimilor $L^p[a,b]$ şi $L^p(R)$. Pentru moment subliniem doar că $L^p[a,b]$ şi $L^p(R)$ sînt spaţii vectoriale peste R în raport cu operaţiile

$$\tilde{f} + \tilde{g} = \widetilde{f + g} \text{ unde } f \in \tilde{f} \text{ şi } g \in \tilde{g}$$

şi

$$\lambda\tilde{f} = \widetilde{\lambda f} \quad \text{unde} \quad f \in \tilde{f}.$$

§ 3. <u>Spaţii metrice. Teorema de punct fix a lui Banach</u>

Fie X o mulţime nevidă.

<u>Definiţia 3.1</u>. O funcţie $d: X \times X \rightarrow R_+$ se numeşte <u>semi-metrică</u> pe X dacă d satisface

SM_1) $d(x,x) = 0$, $(\forall)\ x \in X$

SM_2) $d(x,y) = d(y,x)$, $(\forall)\ (x,y) \in X \times X$

SM_3) $d(x,z) \leq d(x,y) + d(y,z)$ $(\forall)\ (x,y), (y,z) \in X \times X$.

Dacă, în plus, d satisface

M_1) $d(x,y) = 0$ dacă şi numai dacă $x = y$,

d se numeşte <u>metrică</u> sau <u>distanţă</u> pe X.

Cuplul (X,d) în care d este o semi-metrică (metrică) pe X se numeşte spaţiu <u>semi-metric</u> (<u>metric</u>).

<u>Exemple</u>. a) Fie $X = R$ şi $d: R \times R \rightarrow R_+$ definită prin $d(x,y) = |x-y|$, $(\forall)\ (x,y) \in R \times R$. Se verifică cu uşurinţă că (R,d) este un spaţiu metric. Metrica de definită ca mai sus se

numește <u>metrica uzuală</u> pe R. b) Fie $n \in N$, $X = R^n$ și $p \geq 1$. Definim $d_p: R^n \times R^n \to R_+$ prin

$$d_p(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}, \ (\forall)\ (x,y) \in R^n \times R^n,$$

$x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n)$. Evident d_p satisface (SM_1) și (SM_2). In plus, din Inegalitatea lui Minkowski - vezi (1.5) - avem

$$d_p(x,z) = \left(\sum_{i=1}^{n} |x_i - z_i|^p\right)^{1/p} = \left(\sum_{i=1}^{n} |(x_i - y_i) + (y_i - z_i)|^p\right)^{1/p}$$

$$\leq \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i - z_i|^p\right)^{1/p} = d_p(x,y) + d_p(y,z)$$

Deci (R^n, d_p) este un spațiu metric. Facem observația că pentru $p = 2$, această metrică poartă numele de <u>metrica euclidiană</u> pe R^n. Lăsăm în seama cititorului să verifice că $d_\infty : R^n \times R^n \to R_+$ definită prin

$$d_\infty(x,y) = \max\{|x_1 - y_1|, |x_2 - y_2|, \dots, |x_n - y_n|\}$$

$(\forall)\ (x,y) \in R^n \times R^n$ este de asemenea o metrică pe R^n și că

$$d_\infty(x,y) = \lim_{p \to +\infty} d_p(x,y) \quad (\forall)\ (x,y) \in R^n \times R^n.$$

c) Fie $X = C[a,b]$ - mulțimea tuturor funcțiilor continue definite pe $[a,b]$ au valori reale. Fie $d: C[a,b] \times C[a,b] \to R_+$ definită prin

$$d(f,g) = \sup\{|f(t) - g(t)|;\ \ t \in [a,b]\}$$

pentru orice $(f,g) \in C[a,b] \times C[a,b]$. Se constată cu ușurință că d verifică (SM_1) și (SM_2). Pentru a arăta că d verifică (SM_3) să observăm că funcția $|f-g|$ fiind continuă pe $[a,b]$ - care este compact - există $t_o \in [a,b]$ astfel încît

$$d(f,g) = |f(t_o) - g(t_o)| \leq |f(t_o) - h(t_o)| + |h(t_o) - g(t_o)| \leq$$

$$\leq \sup\{|f(t)-h(t)|;\ t\in[a,b]\}+\sup\{|h(t)-g(t)|;\ t\in[a,b]\}=$$
$$= d(f,h) + d(h,g)$$

pentru orice $f,g,h\in C[a,b]$. Deci $(C[a,b],d)$ este un spaţiu metric.

d) Fie $X = \mathcal{L}^p[a,b]$ - mulţimea tuturor funcţiilor definite pe $[a,b]$ cu valori reale cu proprietatea că f şi $|f|^p$ sînt integralele Lebesgue pe $[a,b]$. Aici $p\geq 1$ este un număr fixat. Definim $d_p:\mathcal{L}^p[a,b]\times\mathcal{L}^p[a,b]\ \ R_+$ prin

$$d_p(f,g) = \left(\int_a^b |f(t)-g(t)|^p dt\right)^{1/p}$$

pentru orice $(f,g)\in\mathcal{L}^p[a,b]\times\mathcal{L}^p[a,b]$. Evident d_p verifică (SM_1) şi (SM_2). Pentru a demonstra că d_p verifică şi (SM_3) se procedeaza la fel ca în Exerciţiul 4 din paragraful 7. Facem remarca deosebit de utilă că din $d_p(f,g) = o$ rezultă $f(t) = g(t)$ a.p.t. $t\in[a,b]$. Vezi Consecinţa 2.1. Deci $(\mathcal{L}^p[a,b],\ d_p)$ este un spaţiu semi-metric dar nu este un spaţiu metric.

e) Fie acum $L^p[a,b]$ - mulţimea claselor de echivalenţă - în raport cu egalitatea a.p.t. - de funcţii din $\mathcal{L}^p[a,b]$, unde $p\geq 1$ este un număr fixat. Definim $d_p:\ L^p[a,b]\times L^p[a,b]\rightarrow R_+$ prin

$$d_p(\tilde{f},\tilde{g}) = \left(\int_a^b |f(t)-g(t)|^p dt\right)^{1/p}$$

pentru orice $(\tilde{f},\tilde{g})\in L^p[a,b]\times L^p[a,b]$, unde $f\in\tilde{f}$ şi $g\in\tilde{g}$ sînt reprezentanţi arbitrari. Se observă că d_p nu depinde de alegerea reprezentanţilor $f\in\tilde{f}$ şi $g\in\tilde{g}$. Altfel spus, dacă $f,f_1\in\tilde{f}$ şi $g,g_1\in\tilde{g}$ atunci

$$\left(\int_a^b |f(t)-g(t)|^p dt\right)^{1/p} = \left(\int_a^b |f_1(t)-g_1(t)|^p dt\right)^{1/p}$$

întrucît $f-g = f_1-g_1$ a.p.t. $t\in[a,b]$. Folosind cele stabilite la Exemplul (d), urmează că d_p este o metrică pe $L^p[a,b]$. Deci $(L^p[a,b],d_p)$ este un spaţiu metric.

Definiția 3.2. Un șir (x_n) dintr-un spațiu metric (X,d) se numește

(i) convergent la $x\in X$ dacă $\lim_{n\to+\infty} d(x_n,x) = 0$. In acest caz vom scrie $\lim_{n\to+\infty} x_n = x$, iar x se va numi limita lui (x_n)

(ii) șir Cauchy sau fundamental dacă pentru orice $\varepsilon > 0$ există $n(\varepsilon)\in N$ astfel încît pentru orice $n\in N$, $n\geq n(\varepsilon)$ și orice $p\in N$ să aibă loc

$$d(x_{n+p}, x_n)\leq \varepsilon$$

Propoziția 3.1. Fie (x_n) un șir din (X,d) și $x\in X$. Atunci $\lim_{n\to+\infty} x_n = x$ dacă și numai dacă pentru orice $\varepsilon>0$ există $n(\varepsilon)\in N$ astfel încît pentru orice $n\in N$, $n\geq n(\varepsilon)$ să aibă loc

$$d(x_n,x)\leq \varepsilon$$

Demonstrația fiind foarte simplă o lăsăm pe seama cititorului.

Propoziția 3.2. Dacă $\lim_{n\to+\infty} x_n = x$ și $\lim_{n\to+\infty} x_n = y$ atunci $x=y$. Cu alte cuvinte, limita unui șir convergent dintr-un spațiu metric este unică.

Demonstrație. Din (SM_3) și (SM_2) avem

$$d(x,y)\leq d(x,x_n)+ d(x_n,y) = d(x_n,x)+ d(x_n,y)$$

pentru orice $n\in N$. Trecînd la limită pentru $n\to+\infty$ obținem

$$d(x,y) = 0 ,$$

ceea ce, conform (M_1), implică $x = y$.

Propoziția 3.3. Dacă $\lim_{n\to+\infty} x_n = x$ atunci (x_n) este șir Cauchy.

Demonstrație. Fie $\varepsilon> 0$. Din (SM_3) deducem

$$d(x_{n+p},x_n)\leq d(x_{n+p},x)+ d(x,x_n)$$

pentru orice $n,p\in N$. Din (SM_2) și Propoziția 3.1 urmează că pentru $\varepsilon>0$ există $n(\varepsilon)\in N$ astfel încît pentru orice $n\geq n(\varepsilon)$ să

aibă loc

$$d(x_n,x) \leq \varepsilon/2$$

şi în consecinţă

$$d(x_{n+p},x_n) \leq d(x_{n+p},x) + d(x_n,x) \leq \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Deci (x_n) este şir Cauchy.

Observaţia 3.1. Facem remarca deosebit de utilă că nu pentru orice spaţiu metric (X,d) are loc şi reciproca Propoziţiei 3.3. De exemplu, fie $X = Q$ şi fie $d : Q \times Q \longrightarrow R_+$ definită prin $d(x,y) = |x-y|$ pentru orice $(x,y) \in Q \times Q$. Evident (Q,d) este un spaţiu metric. Pentru orice $n \in N$ să definim $x_n = 1,a_1,a_2,\ldots,a_n$, unde $a_1,a_2,\ldots,a_n$ sînt primele cifre din dezvoltarea zecimală a numărului $\sqrt{2}$.

Să observăm că

$$d(x_{n+p},x_n) = \underbrace{0,00\ldots 0}_{n-\text{ori}}a_{n+1}\,a_{n+2}\ldots a_{n+p} \leq 10^{-n}$$

pentru orice $n \in N$ şi orice $p \in N$. In consecinţă (x_n) este şir Cauchy. Pe de altă parte avem $\lim_{n\to+\infty} |x_n - \sqrt{2}| = 0$ şi atunci $\lim_{n\to+\infty} x_n = \sqrt{2}$ în spaţiul metric (R,d) din Exemplul a.

Dacă (x_n) ar fi convergent în (Q,d) la x, atunci (x_n) ar converge şi în (R,d) la x şi din Propoziţia 3.2 ar rezulta $x = \sqrt{2} \in Q$. Absurditatea la care am ajuns probează că (x_n) nu este convergent în (Q,d).

Această observaţie ne sugerează următoarea

Definiţia 3.3. Un spaţiu metric (X,d) se numeşte spaţiu metric complet dacă orice şir Cauchy din X este convergent.

Cu alte cuvinte, (X,d) este spaţiu metric complet dacă pentru orice şir Cauchy (x_n) din X există $x \in X$ astfel încît $\lim_{n\to+\infty} x_n = x$.

Observaţia 3.2. Subliniem că spaţiile metrice din Exemplele (a), (b), (c) şi (e) sînt spaţii metrice complete.

Definiţia 3.4. Fie (X,d) şi (Y,δ) două spaţii metrice şi fie $f : X \to Y$. Funcţia f este continuă în $x \in X$ dacă pentru orice şir (x_n) din X cu $\lim_{n\to+\infty} x_n = x$ în (X,d) avem $\lim_{n\to+\infty} f(x_n) = f(x)$ în (Y,δ). O funcţie $f : X \to Y$ continuă în orice $x \in X$ se numeşte funcţie continuă pe X.

Definiţia 3.5. Fie (X,d) un spaţiu metric. O funcţie $f : X \to X$ se numeşte contracţie de constantă $k \in (0,1)$ dacă

$$d(f(x), f(y)) \leq kd(x,y) \tag{3.1}$$

pentru orice $(x,y) \in X \times X$.

Propoziţia 3.4. Dacă $f : X \to X$ este o contracţie de constantă $k \in (0,1)$ atunci f este continuă pe X.

Demonstraţie. Evident

$$d(f(x_n), f(x)) \leq kd(x_n,x)$$

pentru orice $x, x_n \in X$. Dacă (x_n) este un şir cu $\lim_{n\to+\infty} x_n = x$, din inegalitatea de mai sus deducem $\lim_{n\to+\infty} f(x_n) = f(x)$, ceea ce încheie demonstraţia.

Definiţia 3.6. Fie $f : X \to X$. Un element $x \in X$ se numeşte punct fix al funcţiei f dacă

$$f(x) = x.$$

După cum vom vedea în capitolele următoare numeroase tipuri de ecuaţii pot fi rescrise echivalent sub forma $f(x) = x$. De aceea este deosebit de important de a pune în evidenţă condiţii suficiente pentru ca o funcţie $f : X \to X$ să admită un punct fix - eventual unic.

Un rezultat fundamental în această privinţă este:

Teorema 3.1. (de punct fix a lui Banach). Fie (X,d) un spaţiu metric complet şi $f : X \to X$ o contracţie de constantă $k \in (0,1)$. Atunci f are un punct fix unic $x \in X$ care este limita în (X,d) a

șirului de aproximații succesive (x_n) definit prin $x_o \in X$ - arbitrar, $x_1 = f(x_o)$, $x_2 = f(x_1), \ldots, x_n = f(x_{n-1}), \ldots$. In plus avem

$$(3.2) \qquad d(x_n, x) \leq \frac{k^n}{1-k} d(f(x_o), x_o)$$

pentru orice $n \in N$.

Observația 3.3. Remarcăm că formula (3.2) ne oferă o evaluare a „vitezei de convergență" a șirului (x_n) la punctul fix al lui f. Această evaluare este deosebit de utilă în procesul de rezolvare aproximativă a ecuațiilor de forma $f(x) = x$, ea permițînd stabilirea numărului minim n de iterații ce trebuie făcute pentru ca iterata de ordin n - x_n - să difere cu mai puțin decît un $\varepsilon > o$ - prescris - de soluția „exactă" x.

Demonstrația Teoremei 3.1. Fie $x_o \in X$ un element arbitrar, care va fi fixat în cele ce urmează și fie (x_n) șirul definit prin $x_1 = f(x_o)$, $x_2 = f(x_1) \ldots$ în general

$$(3.3) \qquad x_n = f(x_{n-1}) \quad \text{pentru} \quad n \in N \quad .$$

Vom arăta că (x_n) este șir Cauchy. Cum (X,d) este spațiu metric complet va rezulta atunci că există $x \in X$ cu $\lim\limits_{n \to +} x_n = x$. Trecînd la limită pentru $n \to +\infty$ în (3.3) și ținînd cont de Propoziția 3.4 vom obține $x = f(x)$. In sfîrșit, vom arăta că x este singurul punct fix al lui f.

Pentru început să observăm că

$$(3.4) \qquad d(x_m, x_{m+1}) \leq k^m d(f(x_o), x_o)$$

pentru orice $m \in N$. Intr-adevăr, aplicînd (3.1) succesiv de m ori obținem

$$\begin{aligned} d(x_m, x_{m+1}) &= d(f(x_{m-1}), f(x_m)) \leq k\, d(x_{m-1}, x_m) = \\ &= k\, d(f(x_{m-2}), f(x_{m-1})) \leq k^2 d(x_{m-2}, x_{m-1}) \leq \ldots \\ &\leq k^m d(x_o, x_1) = k^m d(x_1, x_o) = k^m d(f(x_o), x_o), \end{aligned}$$

pentru orice $m \in N$, ceea ce demonstrează (3.4).

Din (SM_3) obținem

$$d(x_{n+p},x_n) \leq d(x_{n+p},x_{n+p-1})+ d(x_{n+p-1},x_{n+p-2})+\ldots+ d(x_{n+1},x_n)$$

pentru orice $n \in N$ și orice $p \in N$. Din (3.4) avem atunci

$$d(x_{m+p},x_n) \leq [k^n+ k^{n+1}+\ldots+ k^{n+p-1}]d(f(x_o),x_o)$$

pentru orice $n \in N$ și orice $p \in N$, sau echivalent

$$d(x_{n+p},x_n) \leq k^n \frac{1 - k^p}{1 - k} d(f(x_o),x_o)$$

pentru orice $n \in N$ și orice $p \in N$. Din ultima inegalitate ținînd cont că $1 - k^p < 1$ (reamintim că numărul $k \in (o,1)$) obținem

$$(3.4) \qquad d(x_{n+p},x_n) < \frac{k^n}{1-k} d(f(x_o),x_o)$$

pentru orice $n \in N$ și $p \in N$.

Intrucît $k \in (o,1)$ avem $\lim\limits_{n \to +\infty} k^n = o$ și atunci pentru orice $\varepsilon > o$ există $n(\varepsilon) \in N$ astfel încît

$$\frac{k^n}{1-k} d(f(x_o),x_o) \leq \varepsilon$$

pentru orice $n \in N$, $n \geq n(\varepsilon)$. Această inegalitate împreună cu (3.4) demonstrează că (x_n) este șir Cauchy. Cum (X,d) este spațiu metric complet, există $x \in X$ astfel încît

$$\lim_{n \to +\infty} x_n = x.$$

Ținînd cont că $\lim\limits_{n \to +\infty} x_{n-1} = x$, trecînd la limită pentru $n \to +\infty$ în (3.3) și folosind Propoziția 3.4 obținem

$$x = f(x) .$$

Deci x este punct fix al lui f.

Pentru a demonstra că x este singurul punct fix al lui f să presupunem prin reducere la absurd că există $y \in X$ cu $x \neq y$ și

$y = f(y)$. Cum $x \neq y$ din (M_1) avem

$$o \neq d(x,y) = d(f(x),f(y)) \text{ ,}$$

iar din (3.1) obţinem

$$o \neq d(x,y) \leq k\, d(x,y) \text{ .}$$

Impărţind ultima inegalitate prin $d(x,y) > o$, deducem $1 \leq k$, relaţie ce contrazice presupunerea $k \in (o,1)$. Deci $x \in X$ este unicul punct fix al lui f.

Pentru a completa demonstraţia să trecem la limită pentru $p \to +\infty$ în (3.4), ţinînd n fixat.

Obţinem

$$d(x,x_n) \leq \frac{k^n}{1-k}\, d(f(x_o),x_1)$$

pentru orice $n \in N$, inegalitate care în virtutea condiţiei (SM_2) este echivalentă cu (3.2). Cu aceasta demonstraţia este încheiată.

O consecinţă deosebit de utilă în aplicaţii este:

<u>Consecinţa 3.1</u>. <u>Fie</u> (X,d) <u>un spaţiu metric complet şi fie</u> $f : X \to X$ <u>o funcţie cu proprietatea că există</u> $p \in N^*$ <u>astfel încît</u> $f^p = \underbrace{fofo\ldots of}_{p\text{-ori}}$ <u>este o contracţie de constantă</u> $k \in (o,1)$. <u>Atunci</u> f <u>are un punct fix unic</u> $x \in X$ <u>care este limita în</u> (X,d) <u>a şirului de aproximaţii succesive</u> (x_n) <u>definit prin</u> $x_o \in X$ - <u>arbitrar</u> $x_1 = f(x_o)$, $x_2 = f(x_1), \ldots, x_n = f(x_{n-1}), \ldots$. <u>In plus avem</u>

$$d(x_n,x) \leq \frac{k^{n-p}}{1-k}\, d(f^{p+1}(x_o),f^p(x_o)) \tag{3.5}$$

<u>pentru orice</u> $n \in N$, $n \geq p+1$.

<u>Demonstraţie</u>. Din Teorema 3.1 rezultă că funcţia $g = f^p$ are un punct fix unic $x \in X$. Avem atunci

$$f^p(x) = x$$

şi

$$d(x,f(x)) = d(f^p(x),f(f^p(x))) = d(f^p(x),f^p(f(x)))$$
$$\leq k\ d(x,f(x)).$$

Dacă $x \neq f(x)$ atunci $o < d(x,f(x)) \leq k\ d(x,f(x))$ şi deci $1 \leq k$. Absurditatea la care am ajuns ne arată că $d(x,f(x)) = o$ sau că $x = f(x)$. Deci x este punct fix pentru f. Cum orice punct fix al lui f este şi punct fix al lui f^p, iar f^p are numai un singur punct fix, rezultă că f are un punct fix unic.

In sfîrşit, să observăm că (3.5) rezultă din (3.2) aplicată funcţiei f^p, ceea ce încheie demonstraţia.

Observaţia 3.4. Clasa funcţiilor care verifică ipotezele Consecinţei 3.1 este strict mai largă decît cea a funcţiilor care verifică ipotezele Teoremei 3.1. Intr-adevăr, fie (R,d) spaţiul metric din Exemplul (a) şi fie $f : R \to R$ definită prin $f(x)=\cos x$ pentru orice $x \in R$. Presupunînd că f este contracţie de constantă $k \in (o,1)$, rezultă

$$|\cos x - \cos y| \leq k|x-y|$$

pentru orice $x,y \in R$, sau

$$\left|\frac{\cos x - \cos y}{x - y}\right| \leq k$$

pentru orice $x,y \in R$, $x \neq y$. In particular, luînd $x = \frac{\pi}{2}$ şi trecînd la limită pentru $y \to \frac{\pi}{2}$, $y \neq \frac{\pi}{2}$ obţinem

$$|\cos' \tfrac{\pi}{2}| = |\sin \tfrac{\pi}{2}| = 1 \leq k$$

ceea ce contrazice faptul că numărul k este subunitar.

Deci f nu este contracţie. Să observăm însă că din Teorema lui Lagrange avem

$$|f^2(x)-f^2(y)| = |\cos\cos x - \cos\cos y| = |\sin \xi|\,|\cos x - \cos y| =$$
$$= |\sin \xi|\,|\sin \eta|\,|x - y|$$

pentru orice $x,y \in R$, unde $\xi \in (\cos x, \cos y)$ sau

$\xi \in (\cos y, \cos x) \subset [-1,1]$ iar $\theta, \eta \in (x,y)$ sau $\theta, \eta \in (y,x)$. In consecință avem

$$d(f^2(x), f^2(y)) \leq k\, d(x,y)$$

pentru orice $x,y \in R$ unde $k = \sin 1$. Deci f^2 este contracție deși f nu are aceeași proprietate.

§ 4. Spații liniare normate, spații Banach, spații Hilbert

Fie X un spațiu liniar (vectorial) peste R.

Definiția 4.1. O funcție $\|\cdot\| : X \to R_+$ se numește normă pe X dacă ea satisface:

(N_1) $\|x\| = 0$ dacă și numai dacă $x = 0$

(N_2) $\|\lambda x\| = |\lambda| \|x\|$ pentru orice $x \in X$ și orice $\lambda \in R$.

(N_3) $\|x+y\| \leq \|x\| + \|y\|$ pentru orice $x,y \in X$.

Un spațiu liniar X pe care este definită o normă $\|\cdot\|$ se numește spațiu liniar normat.

Exemple. (a). Fie $X = R$ spațiu vectorial peste R și fie $\|\cdot\| : R \to R_+$ definită prin $\|x\| = |x|$ pentru orice $x \in R$. Se verifică imediat că $\|\cdot\|$ este o normă pe R. (b). Fie $X = R^n$ spațiu vectorial peste R și fie $p \geq 1$. Definim $\|\cdot\|_p : R^n \to R_+$ prin

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

pentru orice $x = (x_1, x_2, \ldots, x_n) \in R^n$. Evident $\|\cdot\|_p$ satisface (N_1) și (N_2). Din (1.5) rezultă că $\|\cdot\|_p$ satisface și (N_3) și deci $\|\cdot\|_p$ este o normă pe R^n. In cazul în care $p = 2$ această normă poartă numele de norma euclidiană.

Lăsăm în seama cititorului să verifice că $\|\cdot\|_\infty : R^n \to R_+$ definită prin

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$$

pentru orice $x = (x_1, x_2, \ldots, x_n) \in R^n$ este de asemenea o normă pe R^n și că

$$\|x\|_\infty = \lim_{p\to+\infty} \|x\|_p$$

pentru orice $x \in R^n$. (c) Fie $X = C[a,b]$ - spaţiu liniar al tuturor funcţiilor continue definite pe $[a,b]$ şi cu valori reale, şi fie $\|\cdot\|: C[a,b] \to R_+$ definită prin

$$\|f\| = \sup\{|f(t)|;\ \ t\in[a,b]\}$$

pentru orice $f\in C[a,b]$. Se verifică cu uşurinţă că $\|.\|$ satisface (N_1) şi (N_2). In plus, cum $|f+g|$ este continuă pe $[a,b]$ care este compact există $t_o \in [a,b]$ astfel încît

$$\|f+g\| = |f(t_o) + g(t_o)| \leq |f(t_o)| + |g(t_o)| \leq$$

$$\leq \sup\{|f(t)|;\ t\in[a,b]\} + \sup\{|g(t)|;\ t\in[a,b]\} = \|f\| + \|g\|.$$

Deci $\|.\|$ verifică şi (N_3) şi în consecinţă ea este o normă pe $[a,b]$. Această normă poartă numele de <u>norma Cebîşev</u> sau <u>norma supremum</u> pe $C[a,b]$.

(d) Fie $p\geq 1$ şi fie $X = L^p[a,b]$. Definim $\|.\|: L^p[a,b]\to R_+$ prin

$$\|\tilde{f}\| = \left(\int_a^b |f(t)|^p dt\right)^{1/p}$$

pentru orice $\tilde{f}\in L^p[a,b]$, unde $f\in\tilde{f}$ este un reprezentant arbitrar al clasei $\tilde{f}$. Să observăm că dacă $g\in\tilde{f}$ atunci $f = g$ a.p.t. $t\in[a,b]$ şi atunci

$$\left(\int_a^b |f(t)|^p dt\right)^{1/p} = \left(\int_a^b |g(t)|^p dt\right)^{1/p}.$$

Deci $\|\tilde{f}\|$ este corect definită (nu depinde de alegerea reprezentantului $f\in\tilde{f}$). Evident $\|.\|$ verifică (N_1) şi (N_2). Pentru a arăta că $\|.\|$ verifică şi (N_3) se procedează ca în Exerciţiul 4 din paragraful 7. Deci $\|.\|$ este o normă pe $L^p[a,b]$.

<u>Propoziţia 4.1.</u> <u>Fie</u> X <u>un spaţiu liniar şi</u> $\|.\|: X\to R_+$ <u>o normă pe</u> X. <u>Atunci funcţia</u> $d : X \times X \to R_+$ <u>definită prin</u>

$$d(x,y) = \|x-y\|$$

pentru orice $(x,y) \in X \times X$ este o metrică pe X.

Demonstrația fiind imediată o lăsăm în seama cititorului.

Definiția 4.2. Fie X un spațiu liniar normat cu norma $\|\cdot\|: X \to R_+$. Metrica $d: X \times X \to R_+$ definită ca în Propoziția 4.1 poartă numele de metrica indusă de normă.

Evident, conform Propoziției 4.1, orice spațiu liniar normat este un spațiu metric.

Definiția 4.3. Un spațiu liniar normat care, în raport cu metrica indusă de normă, este spațiu metric complet se numește spațiu Banach.

Observația 4.1. Spațiile liniare normate din Exemplul (a), (b), (c) și (d) sînt toate spații Banach. Un exemplu de spațiu liniar normat care nu este spațiu Banach este următorul. Fie $X = C[0,2]$ și $\|\cdot\|_1: C[0,2] \to R_+$ definită prin

$$\|f\|_1 = \int_0^2 |f(t)|\,dt$$

penzru orice $f \in C[0,2]$. Fie șirul (f_n) din $C[0,2]$ definit prin

$$f_n(t) = \begin{cases} t^n & \text{pentru } t \in [0,1) \\ 1 & \text{pentru } t \in [1,2], \end{cases}$$

pentru orice $n \in N$. Un calcul simplu ne arată că (f_n) este șir Cauchy în raport cu metrica indusă de $\|\cdot\|_1$.

Din Teorema 2.2 rezultă că

$$\lim_{n \to +\infty} \|f_n - f\|_1 = 0 \tag{4.1}$$

unde $f(t) = \lim\limits_{n \to +\infty} f_n(t) = \begin{cases} 0 & \text{dacă } t \in [0,1) \\ 1 & \text{dacă } t \in [1,2]. \end{cases}$

Cum orice g care satisface relația (4.1) este în mod necesar egală a.p.t. cu f (vezi Teorema 2.2) iar orice funcție g cu această proprietate nu este continuă, rezultă că (f_n) nu este convergent în $C[0,2]$ în raport cu metrica indusă de $\|\cdot\|_1$.

<u>Definiția 4.4.</u> Fie X un spațiu liniar. O funcție $<\cdot,\cdot> : X \times X \longrightarrow R_+$ se numește produs scalar dacă ea satisface

(P_1) $<x,x> = o$ dacă şi numai dacă $x = o$

(P_2) $<x,y> = <y,x>$ pentru orice $(x,y) \in X \times X$.

(P_3) $<\alpha x+\beta y,z> = \alpha<x,z> + \beta<y,z>$ pentru orice

$\alpha, \beta \in R$ şi orice $x,y,z \in X$.

Un spațiu liniar dotat cu un produs scalar $<\cdot,\cdot>$ se numește <u>spațiu pre-Hilbert</u>.

<u>Exemple</u> (e) Fie $X = R^n$ şi fie $<\cdot,\cdot> : R^n \times R^n \longrightarrow R_+$ definit prin

$$<x,y> = \sum_{i=1}^{n} x_i y_i$$

pentru orice $x = (x_1,x_2,\ldots,x_n)$ şi orice $y = (y_1,y_2,\ldots,y_n)$ din R^n. Se constată imediat că $<\cdot,\cdot>$ este un produs scalar pe R^n. Acest produs scalar poartă numele de <u>produsul scalar standard</u> pe R^n.

(f) Fie $X = L^2[a,b]$ şi fie $<\cdot,\cdot> : L^2[a,b] \times L^2[a,b] \longrightarrow R_+$ definit prin

$$<\tilde{f},\tilde{g}> = \int_a^b f(t)g(t)dt$$

pentru orice $(f,g) \in L^2[a,b] \times L^2[a,b]$ unde $f \in \tilde{f}$ şi $g \in \tilde{g}$.

Se verifică cu ușurință că $<\cdot,\cdot>$ este un produs scalar pe $L^2[a,b]$.

(g) Fie $\rho : R \longrightarrow R_+$ o funcție continuă cu $\rho(x) > o$ pentru orice $x \in R$ şi fie $\mathcal{L}^2(R,\rho)$ mulțimea tuturor funcțiilor $f: R \longrightarrow R$ cu proprietatea că $f^2\rho$ este integrabilă Lebesgue pe R. Pe $\mathcal{L}^2(R,\rho)$ definim relația r prin $f \, r \, g$ dacă $f(t) = g(t)$ a.p.t. $t \in R$.

Lăsăm în seama cititorului să verifice că r este o relație de echivalență pe $\mathcal{L}^2(R,\rho)$. Mulțimea claselor de echivalență în raport cu r se notează cu $L^2(R,\rho)$. Cum $\mathcal{L}^2(R,\rho)$ este un

spaţiu liniar peste R şi $L^2(R,\rho)$ are aceeaşi proprietate. Pe $L^2(R,\rho)$ definim $<\cdot,\cdot>: L^2(R,\rho) \times L^2(R,\rho) \to R_+$ prin

$$<\tilde{f},\tilde{g}> = \int_{-\infty}^{+\infty} f(t)g(t)\rho(t)dt$$

pentru orice $(\tilde{f},\tilde{g}) \in L^2(R,\rho) \times L^2(R,\rho)$ unde $f \in \tilde{f}$ şi $g \in \tilde{g}$.

Se constată că $<\cdot,\cdot>$ definit ca mai sus este un produs scalar pe $L^2(R,\rho)$. In cazul în care $\rho(t) = 1$ pentru orice $t \in R$ spaţiul $L^2(R,\rho)$ va fi notat cu $L^2(R)$.

Fie X un spaţiu pre-Hilbert cu produsul scalar $<\cdot,\cdot>$. Definim funcţia $\|\cdot\| : X \to R_+$ prin

$$\|x\| = \sqrt{<x,x>} \tag{4.2}$$

pentru orice $x \in X$.

<u>Lema 4.1.</u> <u>(Inegalitatea lui Schwarz).</u> <u>Dacă</u> $<\cdot,\cdot>$ <u>este un produs scalar pe X atunci are loc</u>

$$|<x,y>| \leq \|x\| \|y\| \tag{4.3}$$

<u>pentru orice</u> $x,y \in X$, <u>unde</u> $\|\cdot\|$ <u>este definită prin (4.2).</u>

<u>Demonstraţie.</u> Fie $x,y \in X$ şi să definim funcţia $F : R \to R$ prin

$$F(\lambda) = <x+\lambda y, x+\lambda y>$$

pentru orice $\lambda \in R$. Din Definiţia 4.4 urmează că

$$0 \leq F(\lambda) = <x,x> + 2\lambda <x,y> + \lambda^2 <y,y>$$

pentru orice $\lambda \in R$. Echivalent avem

$$\lambda^2 \|y\|^2 + 2\lambda <x,y> + \|x\|^2 \geq 0$$

pentru orice $\lambda \in R$. In consecinţă

$$[<x,y>]^2 - \|x\|^2 \|y\|^2 \leq 0 ,$$

inegalitate care este echivalentă cu (4.3).

Propoziţia 4.2. Dacă X este un spaţiu pre-Hilbert cu produsul scalar $\langle\cdot,\cdot\rangle$, atunci funcţia $\|\cdot\| : X \to R_+$ definită de (4.2) este o normă pe X (numită norma indusă de produsul scalar, sau normă hilbertiană).

Demonstraţie. Intrucît (N_1) şi (N_2) din Definiţia 4.1 rezultă imediat din Definiţia 4.4, ne vom limita numai la demonstraţia faptului că $\|\cdot\|$ satisface (N_3). Pentru aceasta, fie $x,y \in X$. Avem

$$\|x+y\|^2 = \langle x+y,x+y\rangle = \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle =$$
$$= \|x\|^2 + 2\langle x,y\rangle + \|y\|^2.$$

Din (4.3) obţinem atunci

$$\|x+y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2$$

inegalitate echivalentă cu (N_3).

Definiţia 4.5. Un spaţiu pre-Hilbert care, în raport cu norma indusă de produsul scalar, este spaţiu Banach poartă numele de spaţiu Hilbert.

Observaţia 4.2. Spaţiile pre-Hilbert puse în evidenţă în Exemplele (e), (f) şi (g) sînt toate spaţii Hilbert.

Definiţia 4.6. Fie X un spaţiu Hilbert cu produsul scalar $\langle\cdot,\cdot\rangle$ şi cu norma indusă $\|\cdot\|$. O submulţime F a lui X se numeşte:

(i) ortogonală dacă $\langle x,y\rangle = 0$ pentru orice $x,y \in F$ cu $x \neq y$;

(ii) ortonormată dacă este ortogonală şi $\|x\| = 1$ pentru orice $x \in F$;

(iii) completă dacă este ortonormată şi pentru orice altă submulţime ortonormată G cu $F \subset G$ rezultă $F = G$.

Observaţia 4.3. Dacă F este o submulţime ortogonală care nu conţine vectorul nul, atunci submulţimea $G = \{\|x\|^{-1}x;\ x \in F\}$ este ortonormată.

Dacă (x_n) este un şir din spaţiul Hilbert X atunci seria

$\sum_{n=0}^{+\infty} x_n$ este convergentă la $x \in X$ dacă

$$\lim_{n\to+\infty} \| \sum_{k=0}^{n} x_k - x \| = 0 .$$

Propoziția 4.3. Fie X un spațiu Hilbert cu produsul scalar $<\cdot,\cdot>$ și norma indusă $\|\cdot\|$ și fie $F = \{e_n;\ n \in N\}$ o submulțime ortonormată în X. Atunci pentru orice $x \in X$ avem

$$(4.4) \qquad \sum_{n=0}^{k} |<x,e_n>|^2 \leq \|x\|^2 \quad \text{(inegalitatea lui Bessel)}$$

pentru orice $k \in N$. In consecință seria $\sum_{n=0}^{+\infty} |<x,e_n>|^2$ este convergentă și

$$(4.5) \qquad \sum_{n=0}^{+\infty} |<x,e_n>|^2 \leq \|x\|^2$$

Demonstrație. Avem

$$0 \leq \|x - \sum_{n=0}^{k} <x,e_n>\cdot e_n\|^2 = < x - \sum_{n=0}^{k} <x,e_n>\cdot e_n, x - \sum_{m=0}^{k} <x,e_m> e_m >$$

$$= \|x\|^2 - 2 \sum_{n=0}^{k} |<x,e_n>|^2 + \sum_{n=0}^{k} \sum_{m=0}^{k} <x,e_n> <x,e_m> <e_n,e_m> .$$

Cum $<e_n,e_m> = \delta_{n,m} = \begin{cases} 1 \text{ pentru } n = m \\ 0 \text{ pentru } n \neq m \end{cases}$, obținem

$$0 \leq \|x\|^2 - 2 \sum_{n=0}^{k} |<x,e_n>|^2 + \sum_{n=0}^{k} |<x,e_n>|^2 = \|x\|^2 - \sum_{n=0}^{k} |<x,e_n>|^2$$

pentru orice $k \in N$. Deci are loc (4.4). Trecînd la limită pentru $k \to +\infty$ în (4.4) obținem (4.5).

Următoarea teoremă ne furnizează mai multe condiții necesare și suficiente pentru ca o submulțime ortonormată F să fie completă.

Teorema 4.1. Fie X un spațiu Hilbert cu produsul scalar $<\cdot,\cdot>$ și cu norma indusă $\|\cdot\|$ și fie $F = \{e_n;\ n \in N\} \subset X$ o submulțime ortonormată. Atunci următoarele condiții sînt echivalente:

(4.6) F este completă;

(4.7) din $\langle x,e_n\rangle = 0$ pentru orice $n\in N$ rezultă $x = 0$;

(4.8) pentru orice $x\in X$ avem $x = \sum_{n=0}^{+\infty} \langle x,e_n\rangle\cdot e_n$;

(4.9) pentru orice $x,y\in X$ avem $\langle x,y\rangle = \sum_{n=0}^{+\infty} \langle x,e_n\rangle\cdot\langle y,e_n\rangle$ (identitatea lui Parseval);

(4.10) pentru orice $x\in N$ avem $\|x\|^2 = \sum_{n=0} |\langle x,e_n\rangle|^2$

Demonstraţie. Vom arăta că (4.6) $\Rightarrow$ (4.7) $\Rightarrow$ (4.8) $\Rightarrow$ (4.9) $\Rightarrow$ (4.10) $\Rightarrow$ (4.6).

Incepem cu (4.6) $\Rightarrow$ (4.7). Dacă $\langle x,e_n\rangle = 0$ pentru orice $n\in N$ şi $x\neq 0$ urmează că $y = \|x\|^{-1}x$ are aceeaşi proprietate. In plus $y\notin F$ deoarece $\langle y,y\rangle = 1$. Rezultă atunci că $G = F\cup\{y\}$ este tot o submulţime ortonormată cu $F\subset G$ şi $F\neq G$ fapt care este în contradicţie cu (4.6). Deci (4.6) $\Rightarrow$ (4.7). Pentru a demonstra că (4.7) $\Rightarrow$ (4.8) să observăm pentru început că pentru orice $x\in X$ seria $\sum_{n=0}^{+\infty} \langle x,e_n\rangle\cdot e_n$ este convergentă. Intr-adevăr

$$\Big\|\sum_{n=k}^{k+p} \langle x,e_n\rangle\cdot e_n\Big\|^2 = \Big\langle \sum_{n=k}^{k+p} \langle x,e_n\rangle\cdot e_n, \sum_{m=k}^{k+p} \langle x,e_m\rangle\cdot e_m\Big\rangle$$

$$= \sum_{n=k}^{k+p}\sum_{m=k}^{k+p} \langle x,e_n\rangle\langle x,e_m\rangle\langle e_n,e_m\rangle =$$

$$= \sum_{n=k}^{k+p}\sum_{m=k}^{k+p} \langle x,e_n\rangle\langle x,e_m\rangle\,\delta_{n,m} = \sum_{n=k}^{k+p} |\langle x,e_n\rangle|^2$$

pentru orice $k,p\in N$. Cum seria $\sum_{n=0}^{+\infty} |\langle x,e_n\rangle|^2$ este convergentă (vezi (4.5)) urmează că şirul sumelor parţiale al seriei $\sum_{n=0}^{+\infty} \langle x,e_n\rangle\cdot e_n$ este Cuachy. Dar X este spaţiu Hilbert - deci metric complet - şi ca atare seria $\sum_{n=0}^{+\infty}\langle x,e_n\rangle e_n$ este convergentă.

Fi acum $y = x - \sum_{n=0}^{+\infty} \langle x,e_n\rangle\cdot e_n$. Avem atunci

$$<y,e_k> = <x,e_k> - <\sum_{n=0}^{+\infty} <x,e_n> e_n, e_k> =$$

$$= <x,e_k> - \sum_{n=0}^{+\infty} <x,e_n><e_n,e_k> =$$

$$= <x,e_k> - \sum_{n=0}^{+\infty} <x,e_n> \delta_{n,k} = <x,e_k> - <x,e_k> = 0$$

pentru orice $k \in N$. Deci $y = 0$ sau $x = \sum_{n=0}^{+\infty} <x,e_n> \cdot e_n$.

Pentru a demonstra că (4.8) $\Rightarrow$ (4.9) să observăm că

$$<x,y> = <\sum_{n=0}^{+\infty} <x,e_n> e_n, \sum_{m=0}^{+\infty} <y,e_m> \cdot e_m> =$$

$$= \sum_{n=0}^{+\infty} \sum_{m=0}^{+\infty} <x,e_n><y,e_m><e_n,e_m> =$$

$$= \sum_{n=0}^{+\infty} \sum_{m=0}^{+\infty} <x,e_n><y,e_m> \delta_{n,m} = \sum_{n=0}^{+\infty} <x,e_n><y,e_n> .$$

Vom arăta acum că (4.9) $\Rightarrow$ (4.10). Pentru aceasta este suficient să luăm $x = y$ în (4.9).

In sfîrşit, (4.10) $\Rightarrow$ (4.6). Să presupunem, prin reducere la absurd că deşi are loc (4.10) F nu este completă. Rezultă atunci că există $x \in X$ cu $x \notin F$ şi astfel încît submulţimea $F \cup \{x\}$ să fie ortonormată.

Avem atunci

$$1 = \|x\|^2 = \sum_{n=0}^{+\infty} <x,e_n> \cdot e_n = 0 ,$$

contradicţie care probează că (4.10) $\Rightarrow$ (4.6).

Incheiem acest paragraf cu următoarea teoremă pe care o enunţăm fără însă a o demonstra.

Teorema 4.2. Submulţimea $\left\{\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \sin t, \frac{1}{\sqrt{\pi}} \cos t, \frac{1}{\sqrt{\pi}} \sin 2t, \frac{1}{\sqrt{\pi}} \cos 2t, \ldots, \frac{1}{\sqrt{\pi}} \sin nt, \frac{1}{\sqrt{\pi}} \cos nt, \ldots\right\}$ este completă în $L^2[-\pi,\pi]$ dotat cu produsul scalar $<\cdot,\cdot>$ definit prin

$$\langle\tilde{f},\tilde{g}\rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$$

pentru orice $(\tilde{f},\tilde{g}) \in L^2[-\pi,\pi] \times L^2[-\pi,\pi]$, unde $f \in \tilde{f}$ şi $g \in \tilde{g}$.

§ 5. Operatori liniari şi matrici

Fie X şi Y două spaţii Banach şi fie $U : X \to Y$.

Definiţia 5.1- Funcţia $U : X \to Y$ se numeşte operator liniar sau transformare liniară dacă

$$U(\alpha x + \beta y) = \alpha U(x) + \beta U(y) \tag{5.1}$$

pentru orice $x, y \in X$ şi orice $\alpha, \beta \in R$.

Definiţia 5.2. Un operator liniar $U : X \to Y$ se numeşte continuu în $x \in X$ dacă oricare ar fi şirul (x_n) din X cu $\lim_{n\to+\infty} x_n = x$ avem $\lim_{n\to+\infty} U(x_n) = U(x)$. Un operator liniar care este continuu în orice $x \in X$ se numeşte continuu pe X, sau continuu.

Propoziţia 5.1. Fie $U : X \to Y$ un operator liniar. Atunci U este continuu pe X dacă şi numai dacă este continuu în $x = o$.

Demonstraţie. Evident dacă $U : X \to Y$ este continuu pe X este continuu în $X = o$. Reciproc, să presupunem că U este continuu în $x = o$ şi fie $y \in X$ şi (y_n) un şir din X cu $\lim_{n\to+\infty} y_n = y$. Avem atunci

$\lim_{n\to+\infty} (y_n - y) = o$ şi deci $\lim_{n\to+\infty} U(y_n - y) = o$. In concluzie $\lim_{n\to+\infty} [U(y_n) - U(y)] = o$ sau, echivalent $\lim_{n\to+\infty} U(y_n) = U(y)$. Deci U este continuu pe X.

Propoziţia 5.2. Un operator liniar $U : X \to Y$ este continuu pe X dacă şi numai dacă pentru orice $\varepsilon > o$ există $\delta(\varepsilon) > o$ astfel încît

$$\|U(x)\| \leq \varepsilon$$

pentru orice $x \in X$ cu $\|x\| \leq \delta(\varepsilon)$.

Demonstrație. Evident condiția este suficientă. Vezi Propoziția 5.1. Pentru a demonstra necesitatea să presupunem prin reducere la absurd că deși U este continuu pe X există $\varepsilon_o > o$ astfel încît pentru orice $\delta > o$ să existe cel puțin un element $x_\delta \in X$ cu $\|x_\delta\| \leq \delta$ dar $\|U(x_\delta)\| > \varepsilon_o$. In particular, luînd $\delta_n = \frac{1}{n}$ și notînd x_{δ_n} cu x_n obținem $\|x_n\| \leq \frac{1}{n}$ și $\|U(x_n)\| > \varepsilon_o > o$ pentru orice $n \in N$.

Evident $\lim\limits_{n \to +\infty} x_n = o$, dar $(U(x_n))$ nu converge la $o = U(o)$. Cu aceasta demonstrația este încheiată.

Teorema 5.1. Un operator liniar $U : X \to Y$ este continuu pe X dacă și numai dacă există un număr $\mu > o$ astfel încît

$$\|U(x)\| \leq \mu \|x\| \tag{5.2}$$

pentru orice $x \in X$.

Demonstrație. Din (5.2) și Propoziția 5.2 rezultă că U este continuu pe X. Reciproc, fie $U : X \to Y$ un operator continuu pe X. Din Propoziția 5.2 rezultă că pentru $\varepsilon = 1$ există $\delta > o$ astfel încît

$$\|U(x)\| \leq 1$$

pentru orice $x \in X$ cu $\|x\| \leq \delta$.

In particular, dacă $\lambda > o$ și $x \in X$ cu $\lambda \|x\| \leq \delta$ avem

$$\|U(\lambda x)\| \leq 1$$

sau

$$\|U(x)\| \leq 1/\lambda .$$

Fie acum $x \in X$ cu $x \neq o$ și fie $\lambda = \delta / \|x\|$.

Evident $\lambda \|x\| \leq \delta$ și atunci

$$\|U(x)\| \leq (1/\delta) \|x\|$$

pentru orice $x \in X$ cu $x \neq o$. Luînd $\mu = 1/\delta$ în ultima inegalitate și observînd că aceasta are loc și pentru $x = o$, obținem (5.2).

Fie L(X,Y) mulţimea operatorilor liniari şi continui de la X în Y. Se constată cu uşurinţă că L(X,Y) este un spaţiu liniar peste R în raport cu operaţiile de adunare a doi operatori şi de înmulţire a unui operator cu un scalar.

Teorema 5.2. Funcţia $\|\cdot\|_o : L(X,Y) \longrightarrow R_+$ definită prin

(5.3) $\|U\|_o = \sup\{ \|U(x)\| ; \|x\| \leq 1\}$, pentru orice $U \in L(X,Y)$

este o normă pe L(X,Y) numită norma operatorială.

Demonstraţie. Din (5.1) avem $\|U\|_o \leq \mu$ şi în consecinţă $\|\cdot\|_o$ nu poate lua valoarea $+\infty$.

Evident $\|U\|_o = o \Leftrightarrow U(x) = o$ pentru orice $x \in X$ cu $\|x\| \leq 1$. Dacă $x \in X$, $\|x\| > 1$ atunci $\|x\|^{-1}x$ are norma egală cu 1 şi atunci

$$o = U(\|x\|^{-1}x) = \|x\|^{-1} U(x)$$

ceea ce implică $U(x) = o$. Deci $\|U\|_o = o \Leftrightarrow U = o$ şi atunci $\|\cdot\|_o$ satisface (N_1). Un simplu calcul ne arată că $\|\lambda U\|_o = |\lambda| \|U\|_o$ pentru orice $\lambda \in R$ şi orice $U \in L(X,Y)$. Deci $\|\cdot\|_o$ satisface şi (N_2).

In sfîrşit

$$\|U + V\|_o = \sup\{\|U(x) + V(x)\| ; \|x\| \leq 1\} \leq$$
$$\leq \sup\{\|U(x)\| + \|V(x)\| ; \|x\| \leq 1\} \leq$$
$$\leq \sup\{\|U(x)\| ; \|x\| \leq 1\} + \sup\{\|V(x)\| ; \|x\| \leq 1\} =$$
$$= \|U\|_o + \|V\|_o$$

pentru orice $U,V \in L(X,Y)$. Deci $\|\cdot\|_o$ satisface şi (N_3) şi cu aceasta demonstraţia este încheiată.

Observaţia 5.1. Se poate demonstra că pentru orice $U \in L(X,Y)$ are loc

(5.4) $\|U\|_o = \sup\{\|U(x)\| ; \|x\| = 1\}$.

Vezi de exemplu [17, p.13]. Să mai observăm că $\|U\|_o$ este cel mai mic număr $\mu > o$ care satisface (5.2). Deci

(5.5) $$\|U(x)\| \leq \|U\|_o \|x\|$$

pentru orice $x \in X$.

Teorema 5.3. Fie X, Y şi Z trei spaţii Banach şi fie $U \in L(X,Y)$ şi $V \in L(Y,Z)$. Atunci

(5.6) $$\|V \circ U\|_o \leq \|V\|_o \|U\|_o .$$

Demonstraţie. Se aplică inegalitatea (5.5).

In cele ce urmează vom determina expresia analitică a normei operatoriale în cîteva cazuri remarcabile.

Fie $X = R^n$ şi $Y = R^m$ cu $n \geq 1$, $m \geq 1$ şi fie $U \in L(R^n, R^m)$. Se ştie atunci că într-o pereche de baze $B = \{e_1, e_2, \ldots, e_n\}$ în R^n şi $\{f_1, f_2, \ldots, f_m\}$ în R^m operatorului U îi corespunde matricea

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

unde a_{ij} sînt definiţi prin

$$U(e_j) = \sum_{i=1}^{m} a_{ij} f_i$$

In acest caz avem

$$U(x) = y$$

dacă şi numai dacă $Ax = y$, adică

(5.7) $$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

Observaţia 5.2. Facem remarca deosebit de utilă că pentru orice alegere a normelor pe R^n şi pe R^m orice operator liniar $U : R^n \longrightarrow R^m$ este continuu.

Teorema 5.4. Fie $U\in L(R^n,R^m)$ un operator liniar care în perechea de baze B din R^n B' din R^m acţionează după legea (5.7). In aceste condiţii:

(a) dacă R^n este dotat cu norma $\|\cdot\|_1$ şi R^m cu norma $\|\cdot\|_1$ (vezi Exemplul (a) din § 4) atunci

(5.8) $$\|U\|_o = \max\left\{\sum_{i=1}^{m}|a_{ij}|;\ j = 1,2,\dots,n\right\};$$

(b) dacă R^n este dotat cu norma $\|\cdot\|_\infty$ şi R^m cu norma $\|\cdot\|_\infty$ (vezi Exemplul (a) din § 4) atunci

(5.9) $$U_o = \max\left\{\sum_{j=1}^{n}|a_{ij}|;\ i = 1,2,\dots,m\right\}$$

(c) dacă R^n este dotat cu norma $\|\cdot\|_2$ şi R^m cu norma $\|\cdot\|_2$ (vezi Exemplul (a) din § 4) şi $n = m$, atunci

(5.10) $$\|U\|_o = \sqrt{|\lambda_M|}$$

unde λ_M este valoarea proprie de modul maxim a matricii $A.A'$.

Dacă matricea A este simetrică atunci

(5.11) $$\|U\|_o = |\lambda_M|$$

unde λ_M este valoarea proprie de modul maxim a matricii A.

Demonstraţie. (a) Avem

$$\|U(x)\|_1 = \sum_{i=1}^{m}\left|\sum_{j=1}^{n}a_{ij}\,x_j\right| \leq \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}.x_j| = \sum_{j=1}^{n}|x_j|\left(\sum_{i=1}^{m}|a_{ij}|\right) \leq$$

$$\leq \max\left\{\sum_{i=1}^{m}|a_{ij}|;\ j=1,2,\dots,n\right\}.\|x\|_1 .$$

Să presupunem că $\max\left\{\sum_{i=1}^{m}|a_{ij}|;\ j = 1,2,\dots,n\right\} = \sum_{i=1}^{m}|a_{ik}|$

Atunci luînd $x = (o,o,\dots,o,1,o\dots o)$ obţinem
locul-k

$$\|U(x)\|_1 = \sum_{i=1}^{m}|a_{ik}|$$

Deci $\max\left\{\sum_{i=1}^{m}|a_{ij}|;\ j=1,2,\dots,n\right\}$ este cel mai mic număr μ care satisface (5.2). Din Observaţia (5.1) rezultă (5.8).

(b) Avem

$$\|U(x)\|_{\infty} = \max\left\{\left|\sum_{j=1}^{n} a_{ij}x_j\right|;\ i=1,2,\dots,m\right\} \leq$$

$$\leq \max\left\{\sum_{j=1}^{n}|a_{ij}|;\ i=1,2,\dots,m\right\}\cdot\max\{|x_j|;\ j=1,2,\dots,n\} =$$

$$= \max\left\{\sum_{j=1}^{n}|a_{ij}|;\ i=1,2,\dots,m\right\}\|x\|_{\infty}.$$

Să presupunem că $\max\left\{\sum_{j=1}^{n}|a_{ij}|;\ i=1,2,\dots,m\right\} = \sum_{j=1}^{n}|a_{kj}|.$

Atunci luînd $x = (x_1,x_2,\dots,x_n)$ cu $x_j = \operatorname{sign} a_{kj}$ obţinem

$$\|U(x)\|_{\infty} = \max\left\{\sum_{j=1}^{n}|a_{ij}|;\ i=1,2,\dots,m\right\}.$$

Raţionînd ca şi la punctul (a) obţinem (5.9).

(c) Vom începe cu cazul în care A este simetrică. In acest caz U este autoadjunct şi are n valori proprii reale, nenule $\lambda_1, \lambda_2, \dots, \lambda_n$ cu $|\lambda_1| \geq |\lambda_2| \geq \dots |\lambda_n|$. De asemenea există o bază $\mathcal{B} = \{e_1,e_2,\dots,e_n\}$ în R^n ortonormată conţinînd numai vectori proprii ai lui U, adică

$$U(e_1) = \lambda_i e_i\ ,\quad i = 1,2,\dots,n.$$

Fie $x \in R^n$, $x = \sum_{i=1}^{n} x_i e_i$.

Avem atunci

$$\|U(x)\|_2^2 = \langle U(x), U(x)\rangle = \left\langle U\left(\sum_{i=1}^{n} x_i e_i\right), U\left(\sum_{j=1}^{n} x_j e_j\right)\right\rangle =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i k_j \langle U(e_i), U(e_j)\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \langle \lambda_i e_i, \lambda_j e_j\rangle =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \lambda_i \lambda_j \langle e_i, e_j\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \lambda_i \lambda_j \delta_{ij} =$$

$$= \sum_{i=1}^{n} x_i^2 \lambda_i^2 \leq \lambda_1^2 \sum_{i=1}^{n} x_i^2 .$$

Deci

$$\|U(x)\|_2 \leq |\lambda_1| \, \|x\|_2$$

pentru orice $x \in R^n$. Luînd $x = e_1$ obţinem

$$\|U(e_1)\|_2 = \|\lambda_1 e_1\|_2 = |\lambda_1| \, \|e_1\|_2 = |\lambda_1| .$$

Deci $|\lambda_1|$ este cel mai mic număr pozitiv μ care satisface (5.2). Din Observaţia 5.1 rezultă (5.11).

Pentru a demonstra (5.10) să observăm că A' este matricea operatorului U^* adjunct lui U şi că

$$\|U^*\|_o = \|U\|_o = \sqrt{\|U \circ U^*\|_o} .$$

Cum AA' este matricea operatorului $U \circ U^*$ şi $A \cdot A'$ este simetrică, din (5.11) şi egalitatea de mai sus obţinem (5.10), ceea ce încheie demonstraţia.

<u>Observaţia 5.3</u>. Fie $A = (a_{ij})_{m \times n}$. Notăm cu

(5.12) $$\|A\|_c = \max \left\{ \sum_{i=1}^{m} |a_{ij}| ; \quad j=1,2,\ldots,n \right\}$$

(5.13) $$\|A\|_\ell = \max \left\{ \sum_{j=1}^{n} |a_{ij}| ; \quad i=1,2,\ldots,m \right\}$$

(5.14) $$\|A\|_e = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} .$$

Din Teorema 5.4 rezultă

(5.15) $$\|Ax\|_1 \leq \|A\|_c \, \|x\|_1$$

(5.16) $$\|Ax\|_\infty \leq \|A\|_\ell \, \|x\|_\infty$$

iar un calcul direct conduce la

(5.17) $$\|Ax\|_2 \leq \|A\|_e \, \|x\|_2$$

pentru orice $x \in R^n$.

Numerele $\|A\|_c$, $\|A\|_\ell$ şi $\|A\|_e$ poartă numele de norma matricii A pe coloane, pe linii şi euclidiană. Aceste denumiri se justifică prin faptul că funcţiile $\|\cdot\|_c$, $\|\cdot\|_\ell$ şi $\|\cdot\|_e$ sînt norme pe spaţiul liniar al matricilor de tip m x n cu elemente reale.

§ 6. Elemente de teoria erorilor

Printr-un <u>număr aproximativ</u> înţelegem un număr a care diferă „puţin" de numărul(exact) A.

Dacă $a < A$, a se numeşte <u>aproximaţie prin lipsă</u> a lui A, iar dacă $A < a$, a se numeşte <u>aproximaţie prin adaos</u> a lui A.

<u>Definiţia 6.1.</u> Eroarea numărului aproximativ a este numărul

$$\Delta a = A - a .$$

<u>Eroarea absolută</u> a numărului aproximativ a este numărul

$$\Delta = |A - a| .$$

Intrucît utilizarea numerelor aproximative este dictată de imposibilitatea aflării valorii exacte a lui A, în general <u>nu putem</u> determina nici eroarea, nici eroarea absolută a unui număr aproximativ a. De aceea sîntem conduşi la următoarea definiţie.

<u>Definiţia 6.2.</u> Numim <u>majorant al erorii absolute</u> a numărului aproximativ a orice număr Δ_a care satisface

$$\Delta = |A - a| \leq \Delta_a . \tag{6.1}$$

Evident, din punct de vedere practic este util să se pună în evidenţă un număr <u>cît mai mic</u> Δ_a care să satisfacă (6.1).

<u>Definiţia 6.3.</u> Numim <u>eroare relativă</u> a numărului aproximativ a, numărul

$$\delta = \frac{\Delta}{|A|} = \frac{|A-a|}{|A|} .$$

Numim <u>majorant al erorii relative</u> a numărului aproximativ a orice număr δ_a care satisface

$$\delta \leq \delta_a . \tag{6.2}$$

Observaţia 6.1. Se constată cu uşurinţă că dacă Δ_a este un majorant al erorii absolute atunci

$$\delta_a = \frac{\Delta_a}{a-\Delta_a}$$

este un majorant al erorii relative a numărului aproximativ a.

Prezentăm mai jos principalele cauze care conduc la erori:

(I) Modelarea matematică a problemei de rezolvat. Subliniem că această modelare nu poate descrie perfect fenomenul studiat exceptînd doar un număr foarte mic de situaţii.

(II) Utilizarea unor procedee de trecere la limită.

(III) Utilizarea unor formule conţinînd numere ale căror valori nu pot fi determinate "exact" (π, e, $\sqrt{2}$ etc.).

(IV) Imposibilitatea scrierii sub formă finită a oricărui număr real într-un sistem de numeraţie ales. Această imposibilitate conduce la aşa-numitele erori de rotunjire.

(V). Operarea cu numere aproximative conduce în mod inevitabil la erori numite erori de operare.

Fie $a \in \mathbb{R}_+$. Se ştie că a poate fi scris sub forma zecimală

(6.3) $$a = \alpha_m 10^m + \alpha_{m-1} 10^{m-1} + \ldots + \alpha_{m-n+1} 10^{m-n+1} + \ldots$$

unde $\alpha_i \in \{0,1,2,\ldots,9\}$ iar $\alpha_m \neq 0$.

In anumite situaţii, cînd membrul drept al relaţiei (6.3) este o serie şi nu o sumă finită, sîntem forţaţi să reţinem numai o sumă parţială şi să lucrăm cu un număr aproximativ.

$$b = \beta_m 10^m + \beta_{m-1} 10^{m-1} + \ldots + \beta_{m-n+1} 10^{m-n+1}.$$

Cifrele reţinute din dezvoltarea lui b (β_i cu $i=m, m-1, \ldots, m-n+1$) se numesc cifre semnificative.

Uneori trebuie să introducem un număr oarecare de zerouri la începutul sau la sfîrşitul numărului, cerute de reprezentarea zecimală. De exemplu

$$b = 7.10^{-3} + 0.10^{-4} + 1.10^{-5} + 0.10^{-6} = \underline{0{,}00}7010$$

sau

$$b = 2.10^{9} + 0.10^{8} + 0.10^{7} + 3.10^{6} + 0{,}10^{5} = 2003.\underline{000.000} .$$

Zerourile subliniate mai sus <u>nu</u> sînt cifre semnificative.

<u>Definiția 6.4</u>. Spunem că primele n cifre semnificative ale numărului aproximativ a dat de (6.3) sînt <u>exacte în sens strict</u> dacă eroarea absolută a numărului nu depășește o doime din 10^{m-n+1}, sau cu alte cuvinte dacă $|A-a| \leq \frac{1}{2}\, 10^{m-n+1}$.

Termenul de cifre semnificative exacte în sens strict nu trebuie luat în sens literar. De exemplu, dacă $A = 10$ și $a = 9{,}995$, atunci primele trei cifre semnificative ale lui a sînt exacte în sens strict deși toate sînt diferite de cifrele semnificative corespunzătoare ale lui A.

Dăm fără demonstrație următoarea teoremă:

<u>Teorema 6.1</u>. <u>Dacă numărul aproximativ</u> $a \in R_+$ <u>are n cifre semnificative exacte în sens strict atunci eroarea sa relativă</u> δ <u>satisface</u>

$$\delta \leq \frac{1}{\alpha_m}\, 10^{1-n} . \tag{6.4}$$

Vom stabili în continuare o formulă generală de determinare a erorilor de operare.

Fie $f : R^n \rightarrow R$ o funcție diferențiabilă. Ne propunem să calculăm un număr care să aproximeze

$$u = f(x_1, x_2, \dots, x_n)$$

știind aproximantele $x_i + \Delta x_i$ ale numerelor x_i pentru $i=1,2,\dots,n$.

Avem

$$\Delta u = f(x_1+\Delta x_1,\ x_2+\Delta x_2, \dots, x_n+\Delta x_n) - f(x_1, x_2, \dots, x_n).$$

Dacă numerele $|\Delta x_i|$ sînt suficient de mici, $i=1,2,\dots,n$, putem scrie

$$|\Delta u| \cong |df(x_1,x_2,\dots,x_n)(\Delta x_1,\Delta x_2,\dots,\Delta x_n)| = \left|\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\Delta x_i\right| \leq$$
$$\leq \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right| |\Delta x_i|.$$

Avem atunci

$$|\Delta u| \leq \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right| |\Delta x_i|.$$

Dacă Δ_{x_i} sînt majoranţi ai erorilor absolute pentru $x_i + \Delta x_i$, $i = 1,2,\dots,n$ atunci un majorant al erorii absolute al lui u este

$$\Delta_u = \sum_{i=1}^{n}\left|\frac{\partial f}{\partial x_i}\right| \Delta_{x_i} . \tag{6.5}$$

Din formula precedentă deducem

$$\delta_u = \sum_{i=1}^{n}\left|\frac{\partial}{\partial x_i}(\ln u)\right| \Delta_{x_i} . \tag{6.6}$$

In cazul particular $f(x_1,x_2,\dots,x_n) = \pm x_1 \pm x_2 \pm\dots\pm x_n$ obţinem

$$\Delta_u = \sum_{i=1}^{n} \Delta_{x_i} \tag{6.7}$$

formulă care ne dă un majorant al erorii absolute a unei sume de numere aproximative în funcţie de majoranţii erorilor absolute ale termenilor.

Dacă, în plus, toţi termenii au acelaşi semn, din (6.5) deducem

$$\delta_u = \frac{x_1\delta x_1 + x_2\delta x_2 + \dots + x_n\delta x_n}{x_1 + x_2 + \dots + x_n} \leq \bar{\delta} \tag{6.8}$$

unde $\bar{\delta} = \max\{\delta_{x_1}, \delta_{x_2},\dots,\delta_{x_n}\}$.

Pentru $f(x_1,x_2,\dots,x_n) = x_1x_2,\dots,x_n$, raţionînd ca mai sus obţinem

$$\delta_u = \sum_{i=1}^{n} \delta_{x_i} . \tag{6.9}$$

Incheiem acest paragraf cu cîteva observaţii referitoare la aşa numita <u>problemă inversă în teoria erorilor</u>. Această problemă poare fi enunţată astfel. Care sînt valorile maxime ale majoranţilor erorilor absolute Δ_{x_i} astfel ca eroarea absolută a lui $f(x_1,x_2,\ldots,x_n)$ să nu depăşească un număr pozitiv dat ?

O soluţie a acestei probleme poate fi obţinută pornind de la aşa-numitul principiu al efectelor egale. In virtutea acestui principiu se presupune că toate numerele $\frac{\partial f}{\partial x_i}\Delta_{x_i}$, $i=1,2,\ldots,n$ contribuie în mod egal la eroarea absolută a lui f. Să presupunem deci că intenţionăm să calculăm f cu o eroare absolută mai mică sau egală cu $\Delta_u > 0$.

Presupunînd

$$\left|\frac{\partial f}{\partial x_i}\right|\Delta_{x_i} = \left|\frac{\partial f}{\partial x_j}\right|\Delta_{x_j}$$

pentru orice $i,j = 1,2,\ldots,n$, din (6.5) obţinem

$$\Delta_{x_i} = \frac{\Delta_u}{n\left|\frac{\partial f}{\partial x_i}\right|} \qquad (6.10)$$

pentru $i = 1,2,\ldots,n$, relaţie care ne oferă un majorant al erorii absolute al lui x_i, $i=1,2,\ldots,n$.

<u>Exemplu</u> [7, p.45]. Baza unui cilindru are raza $R = 2m$ şi înălţimea cilindrului este $H = 3m$. Cu ce erori absolute trebuie să determinăm R şi H astfel încît volumul V să poată fi calculat cu o eroare absolută mai mică sau egală cu $0,1\ m^3$?

Avem $V = \pi R^2 H$ şi $\Delta_V = 0,1$. Punînd $R = 2$, $H = 3$ şi $\pi = 3,14$, obţinem

$$\frac{\partial V}{\partial \pi} = R^2 H = 12$$

$$\frac{\partial V}{\partial R} = 2\pi RH = 37,7$$

$$\frac{\partial V}{\partial H} = \pi R^2 = 12,6.$$

Din (6.10) deducem

$$\Delta_{\pi} = \frac{0,1}{3.12} < 0,003$$

$$\Delta_{R} = \frac{0,1}{3.37,7} < 0,001$$

$$\Delta_{H} = \frac{0,1}{3.12,6} < 0,003.$$

Subliniem că aici am considerat V ca o funcţie de trei variabile: π , R şi H , aceasta deoarece nu putem lucra cu valoarea exactă a numărului iraţional π .

§ 7. Exerciţii

1. Să se demonstreze că oricare ar fi p,q $(1,+\infty)$ cu $1/p + 1/q = 1$ şi oricare ar fi şirurile de numere pozitive (x_n) şi (y_n) are loc

$$\sum_{k=1}^{+\infty} x_k y_k \leq \left(\sum_{k=1}^{+\infty} x_k^p\right)^{1/p} \left(\sum_{k=1}^{+\infty} y_k^q\right)^{1/q}.$$

Să se arate că dacă $\sum_{k=1}^{+\infty} x_k^p$ şi $\sum_{k=1}^{+\infty} y_k^q$ sînt convergente atunci şi $\sum_{k=1}^{+\infty} x_k y_k$ este convergentă.

2. Să se demonstreze că oricare ar fi $p \geqslant 1$ şi oricare ar fi şuririle de numere reale (x_n) şi (y_n) are loc

$$\left(\sum_{k=1}^{+\infty} |x_k+y_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^{+\infty} |x_k|^p\right)^{1/p} + \left(\sum_{k=1}^{+\infty} |y_k|^p\right)^{1/p}.$$

Să se arate că dacă $\sum_{k=1}^{+\infty} x_k^p$ şi $\sum_{k=1}^{+\infty} |y_k|^p$ sînt convergente, atunci şi $\sum_{k=1}^{+\infty} |x_k+ y_k|^p$ este convergentă.

3. Să se arate că oricare ar fi $p,q \in (1,+\infty)$ cu $1/p+1/q=1$ şi oricare ar fi $f \in L^p[a,b]$ şi $g \in L^q[a,b]$ are loc

$$\int_a^b f(t)g(t)dt \leq \left(\int_a^b |f(t)|^p dt\right)^{1/p} \left(\int_a^b |g(t)|^q dt\right)^{1/q}.$$

<u>Indicație</u>. Se va ține cont că orice funcție integrabilă Lebesgue este limita a.p.t. a unui șir de funcții etajate ale căror integrale Lebesgue coincid cu integralele Riemann.

4. Să se arate că oricare ar fi $p \geq 1$ și oricare ar fi $f,g \in L^p[a,b]$ are loc

$$\left(\int_a^b |f(t)+g(t)|^p dt\right)^{1/p} \leq \left(\int_a^b |f(t)|^p dt\right)^{1/p} + \left(\int_a^b |g(t)|^p dt\right)^{1/p}.$$

5. Să se arate că orice funcție continuă $f:[a,b] \to R$ este integrabilă Lebesgue și că integrala ei Lebesgue coincide cu integrala ei Riemann.

6. Fie $X \neq \emptyset$ și fie $d_1,d_2,\dots,d_n$ n - metrici pe X.

Să se demonstreze că oricare ar fi numerele pozitive $\lambda_1, \lambda_2,\dots, \lambda_n$ nu toate nule, funcția $d:X \times X \to R_+$ definită prin

$$d(x,y) = \sum_{i=1}^{n} \lambda_i d_i(x,y)$$

pentru orice $(x,y) \in X \times X$, este o metrică pe X.

7. Fie $X = R$ și $d: R \times R \to R_+$ definită prin $d(x,y)=|x-y|$ pentru orice $(x,y) \in R \times R$ și fie $f: R \to R$ o funcție derivabilă cu $\sup\{|f'(x)|;\ x \in R\} = k \in (0,1)$. Să se arate că f este o contracție de constantă k.

8. Fie $X = R^n$ și $d_1: R^n \times R^n \to R_+$ definită prin $d_1(x,y) = \sum_{i=1}^{n} |x_i-y_i|$. Fie A o matrice de tip $n \times n$, $A = (a_{ij})$ cu $\|A\|_c = \max\{\sum_{i=1}^{n} |a_{ij}|;\ j = 1,2,\dots,n\} = k \in (0,1)$ și fie $x^0 \in R^n$. Definim $f: R^n \to R^n$ prin $f(x) = Ax + x^0$ pentru orice $x \in R^n$. Să se demonstreze că f este o contracție de constantă k.

9. Aceeași problemă cînd d_1 se înlocuiește cu d_∞ și $\|A\|_c$ se înlocuiește cu $\|A\|_\ell$.

10. Aceeași problemă cînd d_1 se înlocuiește cu d_2 și $\|A\|_c$ cu $\|A\|_e = \left(\sum_{i,j=1}^{n} a_{ij}^2\right)^{1/2}$.

11. Să se demonstreze că toate spaţiile liniare normate din Exemplele (a), (b), (c) şi (d) - § 4, sînt spaţii Banach.

12. Fie $p \geqslant 1$ şi fie $\ell_p = \{(x_n)_{n\in N};\ x_n \in R,\ (\forall)\ n \in N,\ \sum_{n=0}^{+\infty} |x_n|^p < +\infty\}$.

(a) Să se demonstreze că ℓ_p dotat cu operaţiile de adunare a şirurilor şi de înmulţire a unui şir cu un număr real este un spaţiu liniar peste R.

(b) Fie $\|\cdot\|_p : \ell_p \to R_+$ definită prin

$$\|(x_n)\|_p = \left(\sum_{n=0}^{+\infty} |x_n|^p\right)^{1/p}$$

pentru orice $(x_n) \in \ell_p$. Să se demonstreze cu ℓ_p este un spaţiu Banach în raport cu $\|\cdot\|_p$. Dacă $p = 2$ să se demonstreze că ℓ_2 este spaţiu Hilbert.

13. Să se determine $\|A\|_c$, $\|A\|_\ell$ şi $\|A\|_e$ ştiind că

a) $A = \begin{pmatrix} 1 & -1 & 2 \\ 0 & 4 & 1 \\ 2 & -1 & 0 \end{pmatrix}$ b) $A = \begin{pmatrix} 0,1 & 0,2 & -0,1 \\ 0,9 & 0 & 0 \\ 0,3 & -0,1 & 0,4 \end{pmatrix}$

c) $A = \begin{pmatrix} -0,2 & 03 & 0,6 \\ 0,1 & -0,3 & 0 \\ 0,3 & 0,1 & -0,2 \end{pmatrix}$ d) $A = \begin{pmatrix} 1 & 2 & -4 & 3 \\ 0 & 1 & 0 & -2 \\ -1 & -1 & 3 & 5 \end{pmatrix}$.

14. Care sînt cifrele semnificative ale numerelor

a) $= 5.10^6 + 0.10^5 + 3.10^3 + 4.10 + 0.10^{-1} + 1.10^{-2} + 0,10^{-3} = 5003040,010$

b) $= 2.10^5 + 0.10^4 + 3.10^3 + 0,10^2 = 203000$

c) $= 3.10^{-3} + 2.10^{-4} + 0.10^{-5} = 0,00320$

15. Să se determine $a = \sqrt{2,01} - \sqrt{2}$ cu trei cifre semnificative exacte în sens strict.

16. Să se determine $a = \sqrt{3} - \sqrt{2,04}$ cu patru cifre semnificative exacte în sens strict.

17. Să se determine un majorant al volumului $V = \frac{1}{6}\pi d^3$

al unei sfere cu diametrul $d = 3,7 \pm 0,5$ cm şi $\pi \cong 3,14$.

18. Care sînt majoranţii erorilor absolute pentru $x \cong 15,2$ şi $y \cong 57^0$ pentru ca eroarea absolută a lui $u = 6\,x^2(\log_{10}x - \sin 2y)$ să nu depăşească 10^{-22}?

Capitolul II

IMBUNATATIREA VITEZEI DE CONVERGENTA A SERIILOR

§1. Aproximarea sumelor seriilor numerice convergente

Reamintim că seria numerică

(1.1) $$a_1 + a_2 + \dots + a_n + \dots$$

cu $a_n \in R$ se numeşte **convergentă** dacă şirul sumelor parţiale asociate $\{S_n\}_{n \in N^*}$ (cu $S_n = \sum_{k=1}^{n} a_k$) este convergent. Dacă

(1.2) $$S = \lim_{n \to \infty} S_n$$

numărul S se numeşte suma seriei şi notăm această situaţie prin simbolul $S = \sum_{n=1}^{\infty} a_n$.

Deoarece în general nu este posibil să determinăm suma S a unei serii convergente, apare problema determinării unui număr care să aproximeze cu o eroare prescrisă suma S a seriei date.

Aceasta este problema de care ne vom ocupa în acest paragraf.

Pentru $n \in N^*$ fixat, seria

(1.3) $$\sum_{k=n+1}^{\infty} a_k = a_{n+1} + a_{n+2} + \dots$$

este convergentă, deoarece sumele parţiale $S'_{p-n} = \sum_{k=n+1}^{p} a_k$ sînt egale cu $S'_{p-n} = S_p - S_n$ şi, pentru $p \to \infty$, $S'_{p-n} \to S - S_n$. Dacă notăm $R_n = \sum_{k=n+1}^{\infty} a_k$, atunci, pentru fiecare $n \in N^*$ are loc relaţia

(1.4) $$S = S_n + R_n.$$

Relaţia (1.4) ne asigură că şirul $\{R_n\}_{n \in N^*}$ este convergent la zero cînd $n \to \infty$. Rezultă că oricare ar fi $\varepsilon_1 > 0$, există $n_1 = n_1(\varepsilon_1)$ astfel încît, pentru orice $n > n_1$

(1.5) $$|S - S_n| = |R_n| \leq \varepsilon_1.$$

Relaţia (1.5) ne spune că S_n aproximează suma S cu o eroare $\leq \varepsilon_1$. In situaţiile practice, nici suma S_n nu se poate determina exact deoarece termenii a_k pot avea în reprezentarea zecimală o infinitate de cifre. In acest caz numerele a_k se înlocuiesc prin numerele $\bar{a}_k$ care au un număr finit de zecimale. Numerele $\bar{a}_k$ se ob-

ţin din a_k prin rotunjire.

Reamintim că rotunjirea la m zecimale se realizează astfel:

Dacă $a_k=q+0,p_1p_2\ldots p_mp_{m+1}\ldots$ $(q\in N;\ 0\le p_i\le 9;\ i\in N^*)$ atunci

$$(1.6)\qquad \bar{a}_k=\begin{cases} q+0,p_1\ldots p_m & 0\le p_{m+1}<5\\ q+0,p_1\ldots p_m+10^{-m} & 5<p_{m+1}\le 9\\ q+0,p_1\ldots p_m+10^{-m} & p_{m+1}=5 \text{ şi } \exists p_n>0 \text{ cu } n\geqslant m+2\\ q+0,p_1\ldots p_m & p_{m+1}=5;p_m=\text{par};p_n=0;n\geqslant m+2\\ q+0,p_1\ldots p_m+10^{-m} & p_{m+1}=5;p_m=\text{impar};p_n=0;n\geqslant m+2. \end{cases}$$

Efectuînd rotunjirea după regula (1.6) se comite eroarea

$$(1.7)\qquad |a_k-\bar{a}_k|\le\frac{5}{10^{m+1}}=\tfrac{1}{2}\cdot 10^{-m}.$$

In acest caz suma $\bar{S}_n=\sum_{k=1}^{n}\bar{a}_k$ aproximează suma S_n cu eroarea:

$$(1.8)\qquad |S_n-\bar{S}_n|\le\frac{n}{2}\cdot 10^{-m}.$$

In sfîrşit, dacă $\bar{S}_n$ este rotunjit la $p<m$ zecimale se obţine numărul $\tilde{S}_n$ care satisface inegalitatea:

$$(1.9)\qquad |\bar{S}_n-\tilde{S}_n|\le\tfrac{1}{2}\cdot 10^{-p}.$$

Numărul $\tilde{S}_n$ $(n>n_1)$ va aproxima suma S cu eroarea

$$|S-\tilde{S}_n|=|(S-S_n)+(S_n-\bar{S}_n)+(\bar{S}_n-\tilde{S}_n)|\le\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon$$

unde $\varepsilon_2=\frac{n}{2}\cdot 10^{-m}$, $\varepsilon_3=\frac{1}{2}\cdot 10^{-p}$.

Din cele de mai sus rezultă următoarea procedură de aproximare a sumei seriei numerice convergente (1.1) cu o eroare $<\varepsilon$:

<u>Procedura 1</u>: 1°. alegem numerele pozitive $\varepsilon_1,\varepsilon_2,\varepsilon_3$ astfel încît $\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon$,

2°. determinăm cel mai mic număr $n\in N$ astfel încît $|R_n|<\varepsilon_1$,

3°. determinăm numărul m de zecimale (cu rotunjire) cu care lucrăm ca fiind cel mai mic număr natural care satisface inegalitatea $\frac{1}{2}\cdot 10^{-m}\le\frac{\varepsilon_2}{n}$ (n fiind determinat la pasul 2°.),

4. determinăm $\bar{a}_k$ $(1\le k\le n)$ şi $\bar{S}_n$,

5. facem rotunjirea finală: $|\bar{S}_n-\tilde{S}_n|\le\varepsilon_3$.

Procedura nu indică în ce mod alegem $\varepsilon_1,\varepsilon_2$ şi ε_3. Se obişnuieşte să se folosească următoarea alegere: $\varepsilon_1=\frac{1}{4}\varepsilon$, $\varepsilon_2=\frac{1}{4}\varepsilon$, $\varepsilon_3=\frac{1}{2}\varepsilon$.

In general, determinarea numărului natural n la pasul 2) este o problemă greu de rezolvat. Există însă cazuri particulare importante în care determinarea lui n se face ușor. Are loc următorul rezultat:

<u>Teorema 1.1</u>: <u>Dacă pentru seria convergentă (1.1) există funcția strict descrescătoare</u> $f: R^+ \longrightarrow R^+$ <u>astfel încît</u> $a_n = f(n)$, <u>atunci</u> $\int_1^{\infty} f(x)dx$ <u>este convergentă și</u>

(1.10) $$\int_{n+1}^{\infty} f(x)dx \le R_n \le \int_{n}^{\infty} f(x)dx.$$

<u>Demonstrație</u>: Funcția f fiind descrescătoare, pentru $n \le x \le n+1$ rezultă $a_n = f(n) \geq f(x) \geq f(n+1) = a_{n+1}$; integrînd pe intervalul $[n, n+1]$ obținem:

(1.11) $$a_{n+1} \le \int_n^{n+1} f(x)dx \le a_n \qquad n \geq 1.$$

Sumînd relațiile (1.11) pentru n de la 1 la k rezultă:

(1.12) $$\sum_{n=2}^{k} a_n \le \int_1^{k} f(x)dx \le \sum_{n=1}^{k} a_n \quad .$$

De aici concluzia că seria (1.1) este convergentă dacă și numai dacă $\int_1^{\infty} f(x)dx$ este convergentă. Seria (1.1) fiind convergentă, rezultă inegalitățile:

(1.13) $$\sum_{n=2}^{\infty} a_n \le \int_1^{\infty} f(x)dx \le \sum_{n=1}^{\infty} a_n \; .$$

Aplicînd acest rezultat seriilor convergente $\sum_{n=N}^{\infty} a_n$ și $\sum_{n=N+1}^{\infty} a_n$ se obțin inegalitățile (1.10).

<u>Exemplul 1</u>: Să evaluăm suma seriei $\sum_{n=1}^{\infty} \frac{1}{n^3+1}$ cu o eroare $< \varepsilon = 10^{-3}$.

1. Alegem : $\varepsilon_1 = \frac{1}{4}\cdot 10^{-3}$, $\varepsilon_2 = \frac{1}{4}\cdot 10^{-3}$, $\varepsilon_3 = \frac{1}{2}\cdot 10^{-3}$.

2. Folosind teorema 1.1 găsim: $|R_n| \le \int_n^{\infty} \frac{dx}{x^3+1} \le \int_n^{\infty} \frac{dx}{x^3} = \frac{1}{2n^2}$. Rezolvînd în $\mathbb{N}$ inegalitatea $\frac{1}{2n^2} \le \frac{1}{4}\cdot 10^{-3}$ rezultă $n \geq 45$; luăm $n_1 = 45$.

3. Cu $\varepsilon_2 = \frac{1}{4}\cdot 10^{-3}$ găsim: $\frac{\varepsilon_2}{45} = \frac{10^{-3}}{4\cdot 45} = \frac{5}{9}\cdot 10^{-5} > \frac{1}{2}\cdot 10^{-5}$. Făcînd rotunjirea la 5 zecimale, rezultă: $|a_k - \bar{a}_k| \le \frac{1}{2}\cdot 10^{-5} < \frac{\varepsilon_2}{45}$.

4. Cei 45 termeni sînt trecuți în ordine în următoarele cinci coloane (de la 1 la 9 în coloana I, ș.a.m.d.):

0,50000	0,00100	0,00015	0,00005	0,00002
0,11111	0,00075	0,00012	0,00004	0,00002
0,03571	0,00058	0,00011	0,00004	0,00002
0,01538	0,00045	0,00009	0,00003	0,00002
0,00794	0,00036	0,00008	0,00003	0,00001
0,00461	0,00030	0,00007	0,00003	0,00001

0,00291	0,00024	0,00006	0,00003	0,00001
0,00195	0,00020	0,00006	0,00002	0,00001
0,00137	0,00017	0,00005	0,00002	0,00001 ,

Sumînd cei 45 de termeni obținem: $\bar{S}_{45} = 0{,}68624$.

5. Rotunjind la miimi găsim: $\tilde{S}_{45} = 0{,}686$.

Numărul 0,686 aproximează suma S a seriei date cu o eroare $< 10^{-3}$.

__Teorema 1.2__: __Dacă pentru seria convergentă (1.1) există funcția $f: R^+ \to R^+$ astfel încît $a_n = f(n)$ și în plus $f'(x) < 0, f'' > 0, \forall x \in R^+$, atunci $\int_1^{\infty} f(t)dt$ este convergentă și__

(1.14) $$\sum_{n=1}^{\infty} a_n \le \sum_{n=1}^{N} a_n + \int_N^{\infty} f(x)dx - \frac{1}{2}\cdot a_{N+1} \le \sum_{n=1}^{N} a_n + \frac{1}{2}(a_N - a_{N+1}).$$

__Demonstrație__: Seria (1.1) fiind convergentă, integrala $\int_1^{\infty} f(x)dx$ va fi convergentă (vezi demonstrația teoremei 1.1). Condiția $f''(x) > 0$ pe R^+ asigură că f' este strict crescătoare pe R^+ și în consecință, pentru $n \le t \le n+1 < c < n+2$ (c fiind dat de formula lui Lagrange aplicată funcției f pe intervalul $[n+1, n+2]$) avem:

$$f'(t) \le f'(n+1) < f'(c) = f(n+2) - f(n+1) = a_{n+2} - a_{n+1} .$$

Integrînd de la x la n+1 ($n \le x \le n+1$) obținem:

$$\int_x^{n+1} f'(t)dt \le (a_{n+2} - a_{n+1})\cdot(n+1-x)$$

adică $f(x) \ge a_{n+1} - (a_{n+2} - a_{n+1})\cdot(n+1-x)$. Prin integrarea acestei inegalități pe intervalul $[n, n+1]$ rezultă:

(1.15) $$\int_n^{n+1} f(x)dx \ge \frac{3}{2}a_{n+1} - \frac{1}{2}\cdot a_{n+2} .$$

Convexitatea funcției f ne asigură că graficul ei se află sub coarda ce trece prin punctele (n, a_n) și $(n+1, a_{n+1})$, adică, pe intervalul $[n, n+1]$, are loc inegalitatea: $f(x) \le a_{n+1} + (a_{n+1} - a_n)(x - (n+1))$. Prin integrarea acestei inegalități se obține:

(1.16) $$\int_n^{n+1} f(x)dx \le \frac{1}{2}(a_n + a_{n+1}).$$

Relațiile (1.15) și (1.16) conduc la:

(1.17) $$\frac{3}{2}\cdot a_{n+1} - \frac{1}{2}\cdot a_{n+2} \le \int_n^{n+1} f(x)dx \le \frac{1}{2}(a_n + a_{n+1}).$$

Inegalitatea (1.17) este echivalentă cu:

(1.18) $$a_{n+1} \le \int_n^{n+1} f(x)dx - \frac{1}{2}a_{n+1} + \frac{1}{2}a_{n+2} \le \frac{1}{2}(a_n + a_{n+2}) .$$

Sumînd de la n=N la ∞ obținem:

(1.19) $$\sum_{n=N+1}^{\infty} a_n \le \int_N^{\infty} f(x)dx - \frac{1}{2}\cdot a_{N+1} \le \sum_{n=N+1}^{\infty} a_n + \frac{1}{2}\cdot(a_N - a_{N+1}) .$$

Adăugînd relației (1.19) suma parțială $S_N = \sum_{n=1}^{N} a_n$ obținem:

(1.20) $\quad \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{N} a_n + \int_N^{\infty} f(x)dx - \frac{1}{2}a_{N+1} \leq \sum_{n=1}^{\infty} a_n + \frac{1}{2}(a_N - a_{N+1})$.

Prin urmare, dacă vom aproxima suma S a seriei prin S':

(1.21) $\quad S \approx S' = \sum_{n=1}^{N} a_n + \int_N^{\infty} f(x)dx - \frac{1}{2} \cdot a_{N+1}$

eroarea comisă va fi cel mult $\frac{1}{2}(a_N - a_{N+1})$. Definim corecția C_N și eroarea E_N prin relațiile:

(1.22) $\quad S = S_N + C_N + E_N \qquad C_N = \int_N^{\infty} f(x)dx - \frac{1}{2}a_{N+1}$.

E_N verifică relația

(1.23) $\quad E_N \leq \frac{1}{2}(a_N - a_{N+1})$

__Exemplul 2__: Seriile convergente $\sum_{n=1}^{\infty} \frac{1}{n^p}$ $(p > 1)$ se numesc p-serii. Fiecării p-serii i se poate asocia funcția $f: R^+ \to R^+$ cu $f(x) = \frac{1}{x^p}$ care satisface condițiile teoremei 1.2. In plus,

$$\frac{1}{2}a_{N+1} + C_N = \int_N^{\infty} \frac{dx}{x^p} = \frac{1}{(p-1)N^{p-1}} \ .$$

Formulele (1.21) - (1.23) conduc la următoarea procedură de aproximare a sumei unei serii $\sum_{n=1}^{\infty} a_n$ care satisface condițiile teoremei 1.2:

__Procedura 2__: 1. alegem numerele pozitive $\varepsilon_1, \varepsilon_2, \varepsilon_3$ astfel încît: $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = \varepsilon$,

2. determinăm cel mai mic număr natural N_1 cu proprietatea

$$\frac{1}{2} \cdot (a_{N_1} - a_{N_1+1}) \leq \varepsilon_1,$$

3. determinăm numărul m de zecimale cu care lucrăm ca fiind cel mai mic număr natural care satisface inegalitatea:

$$\frac{1}{2} \cdot 10^{-m} \leq \frac{\varepsilon_2}{N_1 + 2} \ ,$$

4. determinăm $\bar{a}_k$ $(1 \leq k \leq N_1)$ și $\bar{S}_{N_1}$,

5. determinăm corecția $\bar{C}_{N_1}$ prin: $|\bar{C}_{N_1} - C_{N_1}| \leq \frac{1}{2} \cdot 10^{-m}$

6. determinăm $\bar{S}_1 = \bar{S}_{N_1} + \bar{C}_{N_1}$

7. facem rotunjirea finală : $|\bar{S} - \tilde{S}| \leq \varepsilon_3$.

Numărul $\tilde{S}$ va aproxima suma seriei cu o eroare $\leq \varepsilon$.

__Exemplul 3__: Să evaluăm din nou suma seriei $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$ cu o eroare $\leq 10^{-3}$, folosind procedura a 2-a. Funcția $f(x) = \frac{1}{x^2+1}$ satisface condițiile teoremei 1.2 pentru $x \geq 1$. Procedăm astfel:

1. alegem $\varepsilon_1 = \frac{1}{4} \cdot 10^{-3}$, $\varepsilon_2 = \frac{1}{4} \cdot 10^{-3}$, $\varepsilon_3 = \frac{1}{2} \cdot 10^{-3}$,

2. determinăm cel mai mic număr natural N_1 care verifică inegalitatea

$$\frac{1}{2}\cdot\left(\frac{1}{N_1^3+1}-\frac{1}{(N_1+1)^3+1}\right)<\frac{1}{4}\cdot 10^{-3};$$

se găsește $N_1 = 9$,

3. deoarece $\frac{\varepsilon_2}{11} \geq 3\cdot 10^{-5} > \frac{1}{2}\cdot 10^{-5}$ vom lucra cu 5 zecimale,

4. cu $\bar{a}_k$ $(1 \leq k \leq 9)$ evaluați în exemplul 1 găsim: $\bar{S}_9 = 0{,}68098$,

5. pentru evaluarea corecției $\bar{C}_9$, ținem cont că $\left|\int_9^{\infty}\frac{dx}{x^3+1}-\int_9^{\infty}\frac{dx}{x^3}\right| \leq$ $\leq 3\cdot 10^{-6}$ și luăm $\bar{C}_9 = 0{,}00617 - \frac{1}{2}\cdot 0{,}00100 = 0{,}00567$,

6. obținem $\bar{S} = 0{,}68665$,

7. efectuînd rotunjirea finală găsim: $S \simeq \tilde{S} = 0{,}687$.

Observație: Dacă vom evalua termenii șirului sumelor parțiale asociat seriei divergente $\sum_{n=1}^{\infty}\frac{1}{n}$ cu ajutorul calculatorului, începînd de la un loc încolo șirul va deveni constant. Prin urmare, criteriul de oprire bazat pe stabilizarea unui anumit număr de zecimale nu poate fi folosit totdeauna. In cazul seriilor, nu trebuie început calculul sumelor parțiale înainte de a ne convinge că seria este convergentă. ∎

Pentru fiecare serie cu termeni pozitivi putem alege o procedură specifică de evaluare a sumei cu o eroare prescrisă în legătură cu criteriul care asigură convergența seriei. De exemplu, să presupunem că pentru stabilirea convergenței seriei (1.1) am folosit criteriul raportului. Mai exact, să presupunem că există $r \in (0,1)$ și $N \in \mathbb{N}$ astfel încît

$$\frac{a_{n+1}}{a_n} \leq r < 1 \qquad n \geq N. \tag{1.24}$$

Prin inducție completă rezultă că $a_{N+k} \leq r^k \cdot a_N$. Pentru restul R_N se obține estimarea

$$R_N = \sum_{n=N+1}^{\infty} a_n \leq a_N\,(r + r^2 + \dots\ ) = a_N\cdot\frac{1}{1-r}.$$

Dacă a_N și r sînt mici, restul R_N va fi și el mic. Dacă șirul rapoartelor $\left\{\frac{a_{n+1}}{a_n}\right\}$ este descrescător, atunci (1.24) este satisfăcută pentru $r = \frac{a_{N+1}}{a_N}$ și are loc

$$R_N \leq \frac{a_{N+1}}{1-\frac{a_{N+1}}{a_N}}. \tag{1.25}$$

Partea dreaptă a relației (1.25) poate fi calculată pentru fie-

care N şi sumarea se opreşte cînd ea devine $<\varepsilon$ prescris.

<u>Exerciţiu</u>:Arătaţi că dacă $\lim \sqrt[n]{a_n} < 1$ şi şirul $\{\sqrt[n]{a_n}\}$ este descrescător pentru $n > N$ (cu $a_{N+1} < 1$),atunci se poate folosi următoarea estimare a erorii:

$$R_n \leq \frac{a_{N+1}}{1-\sqrt[N+1]{a_{N+1}}} . \qquad (1.26)$$

Ce criteriu de oprire a sumării se poate folosi în acest caz?*

Să ne ocupăm acum de sumarea seriilor alternate.

Fie $\sum_{n=1}^{\infty} a_n$ o serie alternată,adică o serie pentru care $a_1>0$ şi $a_n \cdot a_{n+1}<0$ ($\forall n \in N$).Conform Criteriului lui Leibniz,această serie este convergentă dacă şirul $|a_n|$ este monoton descrescător, convergent la zero;aproximînd suma seriei prin S_n eroarea comisă este mai mică decît modulul primului termen neglijat.Mai mult decît atît,şirul sumelor parţiale de rang par este monoton crescător şi convergent la suma seriei,în timp ce şirul sumelor parţiale de rang impar este monoton descrescător şi convergent. Prin urmare,dacă $|a_{n+1}|<\varepsilon$,suma S_n aproximează suma seriei cu o eroare $<\varepsilon$.Se obţine astfel un criteriu de oprire a sumării.

In multe situaţii sumarea seriilor alternate convergente se poate reduce la sumarea unei serii cu termeni pozitivi asociate.Intr-adevăr,şirul sumelor parţiale de rang par coincide cu şirul sumelor parţiale ale seriei $\sum_{n=1}^{\infty} b_n$,unde $b_n = a_{2n-1} - a_{2n}$; dacă şirul $|a_n|$ este monoton convergent la zero,seria $\sum_{n=1}^{\infty} b_n$ este o serie cu termeni pozitivi convergentă la suma S a seriei iniţiale.

<u>Exemplul 4</u>:Să analizăm posibilităţile de sumare aproximativă a seriei: $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$.Deoarece şirul $\frac{1}{n}$ este monoton descrescător,seria $\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty}\left(\frac{1}{2n-1}-\frac{1}{2n}\right) = \sum_{n=1}^{\infty}\frac{1}{2n(2n-1)}$ este o serie cu termeni pozitivi.Tinînd cont că

$$\int_N^{\infty} \frac{dx}{(2x-1)\cdot 2x} = \frac{1}{2}\ln\left(1+\frac{1}{2N-1}\right)$$

şi că putem determina cu orice eroare dorim $\ln(1+\frac{1}{2N-1})$, putem aplica procedura a doua pentru determinarea sumei seriei date.Vom av

vea nevoie de n=22 termeni pentru a estima suma seriei cu o eroare $\leq 10^{-4}$, deoarece n=22 verifică condiția

$$\frac{1}{2n}\left[\frac{1}{2n(2n-1)}-\frac{1}{(2n+2)(2n+1)}\right]<\frac{1}{4}\cdot 10^{-4}.$$

Dacă folosim ca criteriu de oprire a sumării condiția $|a_{n+1}|<\frac{1}{4}10^{-4}$, am avea nevoie de $n=4\cdot 10^4$ termeni. Prin urmare, prin trecerea la seria $\sum_{n=1}^{\infty} b_n$ asociată, sumarea se face mult mai rapid.*

Să presupunem că pentru seria alternată $\sum_{n=1}^{\infty}(-1)^{n-1}a_n$ $(a_n>0)$:

(1.27) $$a_n-a_{n+1}>a_{n+1}-a_{n+2} \qquad n=1,2,3,\ldots$$

Atunci

$$\sum_{n=N+1}^{\infty}(-1)^{n-1}a_n=(-1)^N\left[a_{N+1}-a_{N+2}+a_{N+3}-\ldots\right]=(-1)^N\left[\tfrac{1}{2}a_{N+1}+\tfrac{1}{2}(a_{N+1}-a_{N+2})-\tfrac{1}{2}(a_{N+2}-a_{N+3})+\tfrac{1}{2}(a_{N+3}-a_{N+4})-\ldots\right]=$$

$$=\frac{(-1)^N}{2}a_{N+1}+\frac{1}{2}\sum_{n=N+1}^{\infty}(-1)^{n-1}(a_n-a_{n+1}).$$

Ultima serie este alternată deoarece $a_n>a_{n+1}$ și termenii converg la zero în valoare absolută dacă seria inițială este convergentă. Folosind testul obișnuit de apreciere a erorii comise aproximând suma seriei printr-o sumă parțială rezultă:

$$\left|\sum_{n=N+1}^{\infty}(-1)^{n-1}a_n-\frac{(-1)^N}{2}a_{N+1}\right|<\frac{1}{2}\cdot(a_{N+1}-a_{N+2}).$$

In concluzie, dacă adunăm primii N termeni ai seriei alternate convergente și corectăm rezultatul adăugind jumătate din termenul care urmează, eroarea comisă este $\frac{1}{2}\cdot(a_{N+1}-a_{N+2})$.

<u>Exemplul 5</u>: Pentru seria din exemplul 4, folosind observația precedentă, va trebui să folosim n=44 termeni pentru aproximarea sumei cu o eroare $<10^{-3}$, deoarece n=44 verifică inegalitatea: $\frac{1}{2}\left(\frac{1}{n+1}-\frac{1}{n+2}\right)\leq\frac{1}{4}\cdot 10^{-3}$. *

Procedura 2 și rezultatele care urmează se aplică unor clase particulare de serii numerice; este vorba de seriile pentru care termenul general a_n este o funcție de n cu proprietăți speciale de monotonie și convexitate. (serii cu termeni pozitivi) și de serii alternate particulare. Doar procedura 1 se poate aplica pentru orice serie convergentă. In cazul general, procedura 1 nu dă rezultate bune în cazul așa-numitelor serii lent convergente.

Definiţia 1.1: Se spune că seria convergentă $\sum_{n=1}^{\infty} a_n$ este lent convergentă dacă pentru obţinerea sumei sale cu o precizie dorită trebuie luat un număr "mare" de termeni.*

Expresia număr "mare" de termeni o vom accepta cu sensul ei uzual.

Deoarece nu este indicat să se folosească procedura 1 pentru estimarea sumei seriilor numerice lent convergente, s-au căutat alte procedeie cu ajutorul cărora să se estimeze rapid sumele unor astfel de serii . Stabilirea unor astfel de procedeie face obiectul a ceea ce vom numi "îmbunătăţirea vitezei de convergenţă a seriilor". Menţionăm că, în literatura de specialitate, se foloseşte cu aceeaşi semnificaţie expresia "ameliorarea convergenţei seriilor (numerice)".

§2. Imbunătăţirea vitezei de convergenţă a seriilor numerice

Vom prezenta transformarea lui Kummer care este în multe cazuri foarte eficientă pentru evaluarea rapidă a sumei unei serii numerice lent convergente.

Fie seria convergentă (1.1) şi S - suma sa. Se alege o serie convergentă ajutătoare

(2.1) $$\sum_{n=1}^{\infty} b_n \qquad (b_n \neq 0)$$

de sumă B cunoscută şi astfel încît să existe limita

(2.2) $$\lim_{n\to\infty} \frac{a_n}{b_n} = q \neq 0.$$

Au loc egalităţile:

(2.3) $$\sum_{n=1}^{\infty} a_n = q\sum_{n=1}^{\infty} b_n + \sum_{n=1}^{\infty}(a_n - qb_n)$$

sau, echivalent,

(2.3) $$S = qB + \sum_{n=1}^{\infty}(a_n - qb_n).$$

Evaluarea sumei seriei (1.1) se reduce în acest caz la evaluarea sumei seriei:

(2.4) $$\sum_{n=1}^{\infty}(a_n - q\cdot b_n)$$

(evaluare care se va face, în cazurile concrete, tot cu ajutorul

procedurilor indicate anterior). Restul seriei (2.4) se poate pune sub forma:

$$R_N = \sum_{n=N+1}^{\infty}(a_n - qb_n) = \sum_{n=N+1}^{\infty} a_n\left(1 - q\cdot\frac{b_n}{a_n}\right) = \sum_{n=N+1}^{\infty}\varepsilon_n\cdot a_n$$

unde $\varepsilon_n = 1 - q\frac{b_n}{a_n} \longrightarrow 0$ cînd $n \to \infty$.

In general, seria (2.4) converge mai rapid decît seria (1.1). Această afirmaţie se poate justifica pentru anumite clase de serii şi pentru anumite alegeri convenabile ale seriilor (2.1).

O clasă de astfel de serii, este clasa seriilor cu termenul general

$$a_n = \frac{\alpha_0 n^p + \alpha_1 n^{p-1} + \dots + \alpha_p}{\beta_0 n^q + \beta_1 n^{q-1} + \dots + \beta_q} \qquad (n=1,2,3,\dots) \tag{2.5}$$

unde $p, q \in \mathbb{N}$, $\alpha_0 > 0$, $\beta_0 > 0$, $q \geq p+2$. In acest caz $\lim n^2 a_n = c \neq \infty$, lucru pe care îl vom nota scriind $a_n = O\left(\frac{1}{n^2}\right)$. In general, notaţia $a_n = O\left(\frac{1}{n^k}\right)$ va avea semnificaţia: $\lim\limits_{n\to\infty} n^k a_n = c \neq \infty$.

Considerăm seriile convergente

$$S^{(m)} = \sum_{n=1}^{\infty} \frac{1}{n(n+1)\cdots(n+m)} \qquad (m=2,3,\dots). \tag{2.6}$$

Folosind egalitatea

$$\frac{1}{n(n+1)\cdots(n+m)} = \frac{1}{m}\left[\frac{1}{n(n+1)\cdots(n+m-1)} - \frac{1}{(n+1)(n+2)\cdots(n+m)}\right]$$

rezultă

$$S_N^{(m)} = \sum_{n=1}^{N} \frac{1}{n(n+1)\cdots(n+m)} = \frac{1}{m}\left[\frac{1}{1\cdot 2\cdots m} - \frac{1}{(N+1)(N+2)\cdots(N+m)}\right]$$

din care deducem

$$S^{(m)} = \frac{1}{m!\,m}, \tag{2.7}$$

Punem termenul general (2.5) sub forma:

$$a_n = \frac{A_1}{n(n+1)} + \frac{A_2}{n(n+1)(n+2)} + \dots + \frac{A_m}{n(n+1)\cdots(n+m)} + a_n^{(m)} \tag{2.8}$$

unde $A_1, A_2, \dots, A_m$ sînt coeficienţi ce urmează să fie determinaţi în aşa fel încît restul $a_n^{(m)}$ să satisfacă relaţia:

$$a_n^{(m)} = O\left(\frac{1}{n^{2+m}}\right). \tag{2.9}$$

Presupunînd că are loc relaţia (2.8) cu condiţia (2.9), se stabileşte că, în mod necesar, coeficienţii A_i $(i=1,2,\dots,m)$ se pot determina prin formulele:

$$\begin{aligned} A_1 &= \lim_{n\to\infty} n(n+1)\, a_n \\ A_2 &= \lim_{n\to\infty} n(n+1)(n+2)\left[a_n - \frac{A_1}{n(n+1)}\right] \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \end{aligned} \tag{2.10}$$

$$A_m = \lim_{n\to\infty} n(n+1)\dots(n+m)\left[a_n - \frac{A_1}{n(n+1)} - \dots - \frac{A_{m-1}}{n(n+1)\dots(n+m-1)}\right]$$

Rămîne de demonstrat că, pentru A_i $(i=1,2,\dots,m)$ determinaţi prin formulele (2.10), termenii $a_n^{(m)}$ satisfac relaţia

$$a_n^{(m)} = a_n - \sum_{j=1}^{m} \frac{A_j}{n(n+1)\dots(n+j)} = O\left(\frac{1}{n^{2+m}}\right).$$

Demonstraţia o propunem ca exerciţiu cititorului.

Luăm ca serie ajutătoare (2.1) seria

$$B = \sum_{n=1}^{\infty}\left[\frac{A_1}{n(n+1)} + \dots + \frac{A_m}{n(n+1)\dots(n+m)}\right] = \frac{A_1}{1!1} + \frac{A_2}{2!2} + \dots + \frac{A_m}{m!m}. \tag{2.11}$$

Are loc egalitatea:

$$S = \sum_{n=1}^{\infty} a_n = B + \sum_{n=1}^{\infty} a_n^{(m)} . \tag{2.12}$$

<u>Exemplul 1</u>: Să determinăm cu o eroare $\le \varepsilon = 10^{-3}$ suma seriei:

$$\sum_{n=1}^{\infty} \frac{1}{n^2+1} . \tag{2.13}$$

Vom pune: $\dfrac{1}{n^2+1} = \dfrac{A_1}{n(n+1)} + \dfrac{A_2}{n(n+1)(n+2)} + a_n^{(2)}$. Rezultă:

$$A_1 = \lim_{n\to\infty} \frac{n(n+1)}{n^2+1} = 1 \quad , \quad A_2 = \lim_{n\to\infty} n(n+1)(n+2)\left[\frac{1}{n^2+1} - \frac{1}{n(n+1)}\right] = 1.$$

deci $a_n^{(2)} = \dfrac{n-3}{n(n+1)(n+2)(n^2+1)}$. Pentru $N > 3$, din teorema 1.1 rezultă

$$R_N = \sum_{n=N+1}^{\infty} \frac{n-3}{n(n+1)(n+2)(n^2+1)} < \int_N^{\infty} \frac{dx}{x^4} = \frac{1}{3N^3} < \frac{1}{4}\cdot 10^{-3}.$$

Rezultă că pentru evaluarea sumei seriei $\sum_{n=1}^{\infty} a_n^{(2)}$ cu o eroare $\le 10^{-3}$ putem lua $N=12$ termeni. Lucrînd cu patru zecimale rotunjite obţinem:

$$S \approx 1,2500 + (-0,1667) + (-0,0083) + 0 + 0,0005 + 0,0004 + 0,0002 +$$
$$+ 0,0002 + 0,0001 + 0,0001 + 0,0001 + 0,0000 = 1,0766.$$

Este de remarcat faptul că numai 7 din cei 12 termeni necesari evaluării cu eroarea prescrisă au una din cele patru zecimale nenulă. Aceasta se datoreşte faptului că determinarea lui $N=12$ s-a făcut pe baza unei majorări grosiere dar comodă în calcule. Eroarea comisă lucrînd în aceste condiţii este:

$$\frac{1}{4}\cdot 10^{-3} + \frac{9}{2}\, 10^{-4} + 4\cdot 10^{-4} = 1,1\cdot 10^{-3} .$$

Procedînd în acest fel am obţinut : $S = 1,077 \pm 1,1\cdot 10^{-3}$. Pentru a obţine o eroare $< 10^{-3}$ ar fi trebuit să lucrăm cu 5 zecimale. Propunem cititorului să facă acest lucru şi apoi să compare rezultatul găsit cu cel stabilit mai sus şi de asemenea să compa-

re şi volumul de lucru necesar în ambele situaţii.*

Observaţie:1.Procedeul de estimare a sumei seriei din exemplul de mai sus nu se încadrează în nici una dintre procedurile indicate anterior deoarece numărul de zecimale cu care s-a lucrat nu a fost determinat pe baza unei reguli.Acesta este modul de lucru obişnuit al practicienilor care au observat că pentru obţinerea unui rezultat cu o eroare $<10^{-m}$ de obicei este suficient să lucreze cu cel mult m+2 zecimale.Lucrînd astfel apare obligaţia ca,în final,să se estimeze eroarea comisă procedînd în acest fel.

2.Folosirea seriei (2.11) pentru realizarea transformării lui Kummer pentru o serie (2.5) este sugerată de dezvoltarea Stirling a seriei (2.5):

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{p} a_n + A_1 \frac{1}{p+1} + \frac{A_2}{2}\frac{1}{(p+1)(p+2)} + \cdots + \frac{A_m}{m}\frac{1}{(p+1)(p+2)\cdots(p+m)} + \cdots;$$

3.Folosirea transformării lui Kummer presupune cunoaşterea unui set de serii numerice convergente din care să alegem seria ajutătoare (2.1).Un astfel de set se poate găsi în [7].*

§3.Imbunătăţirea vitezei de convergenţă a seriilor întregi prin metoda Euler-Abel

Considerăm seria întreagă convergentă

$$\sum_{n=0}^{\infty} a_n x^n \tag{3.1}$$

avînd suma f(x) pentru $|x|<R$, R fiind finit,diferit de zero. Fără a afecta generalitatea putem presupune că R=1,deoarece,în caz contrar,făcînd tranformarea $t=x/R$ se obţine o serie întreagă în raport cu variabila t şi avînd raza de convergenţă egală cu 1.Punem seria (3.1) sub forma

$$f(x) = a_0 + x\cdot\varphi(x) \tag{3.2}$$

unde

$$\varphi(x) = \sum_{n=1}^{\infty} a_n x^{n-1} = \sum_{n=0}^{\infty} a_{n+1} x^n \quad . \tag{3.3}$$

Rezultă că

$$(1-x)\varphi(x) = a_0 + \sum_{n=0}^{\infty} (a_{n+1} - a_n)\, x^n = a_0 + \sum_{n=0}^{\infty} (\Delta a_n) x^n$$

unde

(3.4)
$$\Delta a_n = a_{n+1} - a_n \qquad (n=0,1,2,\ldots).$$
sînt diferenţele progresive de ordinul întîi ale coeficienţilor a_n. Prin urmare
$$\varphi(x) = \frac{a_0}{1-x} + \frac{1}{1-x}\sum_{n=0}^{\infty}(\Delta a_n)\, x^n$$
şi

(3.5)
$$f(x) = \frac{a_0}{1-x} + \frac{x}{1-x}\sum_{n=0}^{\infty}(\Delta a_n)\cdot x^n \quad ,$$
Seria (3.5) se numeşte <u>transformata Euler-Abel</u> a seriei (3.1). Aplicînd transformata Euler-Abel seriei $\sum_{n=0}^{\infty}(\Delta a_n)\, x^n$ şi folosind notaţia

(3.6)
$$\Delta^2 a_n = \Delta a_{n+1} - \Delta a_n$$
rezultă

(3.7)
$$f(x) = \frac{a_0}{1-x} + \frac{\Delta a_0 \cdot x}{(1-x)^2} + \frac{x^2}{(1-x)^2}\sum_{n=0}^{\infty}(\Delta^2 a_n)\, x^n .$$
Introducînd recursiv diferenţele progresive de ordinul p ale lui a_n prin

(3.8)
$$\Delta^p a_n = \Delta^{p-1} a_{n+1} - \Delta^{p-1} a_n$$
şi folosind inducţia completă se obţine:

(3.9)
$$f(x) = \sum_{k=0}^{p-1}\frac{x^k \cdot \Delta^k a_0}{(1-x)^{k+1}} + \left(\frac{x}{1-x}\right)^p \sum_{n=0}^{\infty}(\Delta^p a_n)\, x^n$$
(cu convenţia $\Delta^0 a_n = a_n$).

<u>Exemplul 1</u>: Să evaluăm cu o eroare $<\varepsilon=10^{-3}$ suma seriei
$$f(x) = \sum_{n=0}^{\infty}\frac{x^n}{(n+1)(n+2)}$$
pentru $x=-1$.

Vom aplica de două ori transformarea Euler-Abel. Au loc egalităţile:
$$a_n = \frac{1}{(n+1)(n+2)} ;\ \Delta a_n = \frac{-2}{(n+1)(n+2)(n+3)} ;\ \Delta^2 a_n = \frac{6}{(n+1)(n+2)(n+3)(n+4)} .$$
Prin urmare $a_0 = \frac{1}{1\cdot 2}$, $\Delta a_0 = -\frac{2}{1\cdot 2\cdot 3}$, iar (3.9) conduce la:
$$f(-1) = \frac{1}{1\cdot 2}\cdot\frac{1}{2} + \frac{2}{1\cdot 2\cdot 3}\cdot\frac{1}{4} + 6\cdot\left(-\frac{1}{2}\right)^2\cdot\sum_{n=0}^{\infty}\frac{(-1)^n}{(n+1)(n+2)(n+3)(n+4)} .$$
Determinăm numărul de termeni pe care îi vom suma din condiţia:
$$\frac{1}{2}\left(\frac{1}{(n+2)(n+3)(n+4)(n+5)} - \frac{1}{(n+3)(n+4)(n+5)(n+6)}\right) < \frac{1}{4}\cdot 10^{-3}$$
şi găsim N=3. Deci adăugînd corecţia $\frac{1}{2}a_4$ sumei parţiale S_3 obţinem estimarea:
$$f(-1) \approx \frac{1}{4} + \frac{1}{12} + \frac{3}{2}\frac{1}{24} - \frac{3}{2}\frac{1}{120} + \frac{3}{2}\frac{1}{360} - \frac{3}{2}\frac{1}{840} .$$
Lucrînd cu 5 zecimale rotunjite obţinem:

$$f(-1) \approx 0,25000+0,08333+0,06250-0,01250+0,00417-0,00090 = 0,38660 .$$

Eroarea pe care o comitem lucrînd în acest fel este:

$$< \frac{1}{4}\cdot 10^{-3} + 3\cdot\frac{1}{4}\cdot 10^{-5} + 6\cdot 10^{-4} = \frac{173}{200}\cdot 10^{-3} < 10^{-3}.$$

Valoarea exactă a lui $f(-1)$ este : $f(-1) = 2\ln 2 - 1 = 0,38630..$

Observaţie: In cazul în care a_n este un polinom de grad $\leq p-1$ în necunoscuta n, $\Delta^p a_n = 0$ şi, din formula (3.9) obţinem:

(3.10) $$f(x) = \sum_{k=0}^{p-1} \frac{x^k \cdot \Delta^k a_0}{(1-x)^{k+1}}$$

§4. Imbunătăţirea vitezei de convergenţă a seriilor Fourier prin metoda lui Krîlov

Fie $f:R \to R$ o funcţie periodică de perioadă $T=2\pi$. Presupunem că f este o funcţie continuă pe porţiuni ale intervalului $[-\pi,\pi]$ şi are derivatele $f^{(i)}(i=1,2,...,m)$ continue pe porţiuni. Aplicînd teorema de convergenţă pentru seria Fourier asociată, rezultă că, în toate punctele de continuitate ale funcţiei f, are loc egalitatea:

(4.1) $$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n\cdot \cos nx + b_n\cdot \sin nx)$$

cu

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \quad , \quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx.$$

In general coeficienţii a_n şi b_n converg lent la zero şi practic este greu de folosit seria (4.1).

Metoda lui Krîlov constă în a extrage din funcţia f o funcţie elementară $g(x)$ (care este în general o funcţie polinomială pe porţiuni) de aceleaşi discontinuităţi ca şi $f(x)$, derivatele $g^{(i)}$ $(i=1,2,..,m)$ avînd aceleaşi discontinuităţi ca derivatele corespunzătoare $f^{(i)}$ ale funcţiei f şi în plus

$$f^{(i)}(-\pi+0)-g^{(i)}(-\pi+0)=f^{(i)}(\pi+0)-g^{(i)}(\pi+0)$$

$(i=0,1,2,..,m)$. In acest caz diferenţa $\varphi(x)= f(x) - g(x)$ are derivatele corespunzătoare (pînă la ordinul m inclusiv) continue pe $[-\pi,\pi]$ şi $\varphi^{(k)}(-\pi+0) = \varphi^{(k)}(\pi-0)$ pentru $k=0,1,2,...,m$, adică valorile funcţiei φ şi ale primelor m derivate coincid în extre-

mitățile intervalului $[-\pi,\pi]$. Se spune în acest caz că φ aparține clasei de periodicitate $\tilde{C}^m$. Vom avea

(4.2) $$f(x) = g(x) + \left[\tfrac{1}{2}\alpha_0 + \sum_{n=1}^{\infty}(\alpha_n \cos nx + \beta_n \sin nx)\right]$$

unde α_n și β_n sînt coeficienții Fourier ai funcției $\varphi(x)$ și satisfac condiția $\lim\limits_{n\to\infty} n^m\alpha_n = \lim\limits_{n\to\infty} n^m\beta_n = 0$, adică seria (4.2) va fi rapid convergentă.

Să construim o funcție $g(x)$ care să satisfacă condițiile prezentate anterior.

Construim șirul de funcții $\sigma_0(x), \sigma_1(x), \dots, \sigma_m(x)$ definite pe intervalul $[-2\pi, 2\pi]$ care au proprietățile:

(4.3) $$\sigma_k^{(k)}(+0) - \sigma_k^{(k)}(-0) = \pi \qquad k=0,1,2,\dots,m,$$

și astfel încît derivatele $\sigma_k^{(j)}(x)$ $(j=0,1,2,\dots,k-1)$ să fie continue pe intervalul $[-2\pi, 2\pi]$. Funcția $\sigma_0(x)$ este definită prin

(4.4) $$\sigma_0(x) = \begin{cases} \dfrac{-\pi - x}{2} & -2\pi < x < 0 \\ \dfrac{\pi - x}{2} & 0 < x < 2\pi \\ 0 & x \in \{-2\pi, 0, 2\pi\}. \end{cases}$$

Rezultă că

(4.5) $$\sigma_0(x) = \sum_{n=1}^{\infty} \frac{\sin nx}{n}$$

$\sigma_0(x)$ are o discontinuitate în $x=0$ cu saltul π, adică:

$$\sigma_0(+0) - \sigma_0(-0) = \tfrac{1}{2}\pi - (-\tfrac{1}{2}\pi) = \pi$$

și de aceea funcția $\psi(x) = \sigma_0(x-x_0)$ are în x_0 același salt ca $\sigma_0(x)$. Definim

(4.6) $$\sigma_1(x) = c_1 + \int_0^x \sigma_0(t)\,dt.$$

Integrînd (4.5) termen cu termen obținem:

(4.7) $$\sigma_1(x) = c_1 + \sum_{n=1}^{\infty}\frac{1}{n^2} - \sum_{n=1}^{\infty}\frac{\cos nx}{n^2}$$

Alegem $c_1 = -\sum\limits_{n=1}^{\infty}\dfrac{1}{n^2} = -\dfrac{\pi^2}{6}$ și găsim

(4.8) $$\sigma_1(x) = -\sum_{n=1}^{\infty}\frac{\cos nx}{n^2}.$$

Integrînd acum (4.4) obținem:

$$\sigma_1(x) = \begin{cases} \dfrac{\pi^2}{12} - \dfrac{(\pi-x)^2}{4} & 0 < x \le 2\pi \\ \dfrac{\pi^2}{12} - \dfrac{(\pi+x)^2}{4} & -2\pi \le x \le 0 \end{cases}$$

adică $\sigma_1(x)$ este continuă pe $[-2\pi, 2\pi]$, dar derivata ei are o

discontinuitate în $x=0$ şi $\sigma_1'(+0) - \sigma_1'(-0) = \pi$. Similar definim

$$\sigma_k(x) = \int_0^x \sigma_{k-1}(t)\,dt + c_k \qquad k=2,3,\dots,m$$

constantele c_k determinîndu-se din condiţia ca termenul liber al seriei Fourier corespunzătoare să fie nul, adică

$$\int_0^{\pi}\left[c_k + \int_0^x \sigma_{k-1}(t)\,dt\right] dx = 0.$$

Funcţiile $\sigma_k(x)$ $(k=1,2,\dots,m)$ şi $\sigma_k^{(j)}(x)$ $(j=1,2,\dots,k-1)$ sînt continue pe $[-2\pi, 2\pi]$ dar $\sigma_k^{(k)}(x) = \sigma_0(x)$ şi are saltul:

$$\sigma_k^{(k)}(+0) - \sigma_k^{(k)}(-0) = \pi \qquad k=1,2,\dots,m.$$

Rezultă că $\psi_k(x) = \sigma_k(x-x_0)$ are proprietatea:

$$\psi_k^{(k)}(x_0+0) - \psi_k^{(k)}(x_0-0) = \pi.$$

Fie acum: $x_1^{(0)}, x_2^{(0)}, \dots, x_{k_0}^{(0)}$ - punctele de discontinuitate ale funcţiei $f(x)$ şi $x_1^{(j)}, x_2^{(j)}, \dots, x_{k_j}^{(j)}$ - punctele de discontinuitate ale funcţiei $f^{(j)}(x)$ $(j=1,2,\dots,m)$. Introducem notaţiile:

$$f^{(p)}(x_j^{(p)}+0) - f^{(p)}(x_j^{(p)}-0) = h_j^{(p)}$$

$(p=1,2,\dots,m\ ; j=1,2,\dots,k_p)$. Definim funcţia $g(x)$ prin formula:

$$g(x) = \sum_{s=1}^{k_0} \frac{h_s^{(0)}}{\pi}\sigma_0(x-x_s^{(0)}) + \sum_{s=1}^{k_1} \frac{h_s^{(1)}}{\pi}\sigma_1(x-x_s^{(1)}) + \dots + \sum_{s=1}^{k_m} \frac{h_s^{(m)}}{\pi}\sigma_m(x-x_s^{(m)}) \qquad (4.9)$$

Funcţia $g(x)$ astfel definită are proprietăţile anunţate la începutul paragrafului. Lăsăm acest lucru în seama cititorului care poate eventual consulta [7], pag.216.

Observaţie: Dacă în extremităţile segmentului $[-\pi, \pi]$ valorile limită ale funcţiei $f(x)$ sau ale derivatelor sale $f^{(k)}$ $(k \le m)$ nu coincid, adică dacă $f^{(p)}(-\pi+0) \neq f^{(p)}(\pi-0)$ $(p=1,2,\dots,k)$ punctele $x=-\pi$ şi $x=\pi$ sînt puncte de discontinuitate. Datorită periodicităţii funcţiei $\sigma_p(x)$, avem $\sigma_p(x+\pi) = \sigma_p(x-\pi)$ şi funcţia $\sigma_p^{(p)}(x+\pi)$ are pe segmentul $[-\pi, \pi]$ două puncte de discontinuitate ($x=\pi$ şi $x=-\pi$) de salt egal cu π. De aceea, în formula (4.9) se include numai contribuţia corespunzătoare unuia dintre punctele $x=\pi$ şi $x=-\pi$.

Exemplul 1: Considerăm funcţia periodică, de perioadă $T=2\pi$ definită pe intervalul $[-\pi, \pi]$ prin:

$$f(x) = \begin{cases} x^2+1 & -\pi < x < 0 \\ x^2 & 0 < x < \pi \end{cases}.$$

Funcţia are punctele de discontinuitate $x_1 = -\pi, x_2 = 0, x_3 = \pi$ în intervalul $[-\pi, \pi]$. Seria Fourier asociată este

$$f(x) = \frac{1}{2} + \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos nx - \frac{2}{\pi}\sum_{k=0}^{\infty} \frac{1}{2k+1} \sin(2k+1)x$$

şi este lent convergentă deoarece $b_n = O(\frac{1}{n})$. Calculăm salturile:

$$h_1^{(0)} = f(-\pi+0) - f(\pi-0) = \pi^2 + 1 - \pi^2 = 1$$

$$h_2^{(0)} = f(+0) - f(-0) = 0 - 1 = -1$$

$$h_3^{(0)} = f(\pi+0) - f(\pi-0) = h_1^{(0)} = 1.$$

Folosind formula (4.9) obţinem:

$$g(x) = \frac{1}{\pi}\sigma_0(x+\pi) - \frac{1}{\pi}\sigma_0(x) = \begin{cases} \frac{1}{2} & -\pi < x < 0 \\ -\frac{1}{2} & 0 < x < \pi \end{cases}$$

Ţinînd cont de formula (4.5) găsim: $g(x) = -\frac{2}{\pi}\sum_{k=0}^{\infty} \frac{\sin(2k+1)x}{2k+1}$.

Prin urmare

$$f(x) = g(x) + \frac{1}{2} + \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \cos nx .$$

§5. Sumarea aproximativă a seriilor trigonometrice

Considerăm seria trigonometrică convergentă

$$(5.1) \qquad \sum_{n=0}^{\infty} (a_n \cos nx + b_n \sin nx)$$

a cărei sumă $S(x)$ este necunoscută. Ne punem problema determinării valorii aproximative a acestei sume cu o precizie prescrisă. Deoarece în general coeficienţii a_n şi b_n converg lent la zero încercăm mai întîi să extragem din seria dată o anumită serie a cărei sumă $g(x)$ este cunoscută şi astfel încît seria rezultantă

$$(5.2) \qquad \sum_{n=0}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx)$$

să aibă o convergenţă mai rapidă decît seria iniţială. Dacă $g(x) = \sum_{n=0}^{\infty} (\bar{a}_n \cos nx + \bar{b}_n \sin nx)$ atunci

$$S(x) = g(x) + \sum_{n=0}^{\infty} (\alpha_n \cos nx + \beta_n \sin nx)$$

cu $\alpha_n = a_n - \bar{a}_n$ şi $\beta_n = b_n - \bar{b}_n$ $(n = 0, 1, 2, \ldots)$.

De obicei drept serii $g(x)$ se folosesc seriile

$$\sigma_0(x) = \frac{\pi - x}{2} = \sum_{n=1}^{\infty} \frac{\sin nx}{n} \qquad (0 < x < 2\pi)$$

$$-\sigma_1(x) = \frac{(\pi-x)^2}{4} - \frac{\pi^2}{12} = \sum_{n=1}^{\infty} \frac{\cos nx}{n^2} \qquad (0 \le x \le 2\pi)$$

$$-\sigma_2(x) = \frac{2\pi^2 x - 3\pi x^2 + x^3}{12} = \sum_{n=1}^{\infty} \frac{\sin nx}{n^3} \qquad (0 \le x \le 2\pi)$$

$$-\ln\left(2 \sin \tfrac{1}{2}x\right) = \sum_{n=1}^{\infty} \frac{\cos nx}{n} \qquad (0 < x < 2\pi)$$

$$-\int_0^x \ln(2\sin\tfrac{1}{2}t)\,dt = \sum_{n=1}^{\infty}\frac{\sin nx}{n^2} \qquad 0\le x\le 2\pi,$$

ş.a.m.d..

Pornind dela observaţia că în seriile de mai sus coeficienţii a_n şi b_n sînt de forma $\frac{1}{n^k}$, pentru serii trigonometrice pentru care a_n şi b_n sînt funcţii raţionale în n cu gradul numărătorului mai mic cel puţin cu o unitate decît gradul numitorului, se utilizează următoarea procedură de aproximare: se scriu coeficienţii a_n şi b_n sub forma

$$a_n = \frac{p_1}{n} + \frac{p_2}{n^2} + \dots + \frac{p_k}{n^k} + \tau_n$$

$$b_n = \frac{q_1}{n} + \frac{q_2}{n^2} + \dots + \frac{q_k}{n^k} + \delta_n$$

şi se obţine

$$S(x) = p_1\sum_{n=1}^{\infty}\frac{\cos nx}{n} + q_1\sum_{n=1}^{\infty}\frac{\sin nx}{n} + \dots + p_k\sum_{n=1}^{\infty}\frac{\cos nx}{n^k} + q_k\sum_{n=1}^{\infty}\frac{\sin nx}{n^k} + \sum_{n=1}^{\infty}(\tau_n\cos nx + \delta_n \sin nx)$$

seriile $\sum_{n=1}^{\infty}\tau_n\cos nx$ şi $\sum_{n=1}^{\infty}\delta_n\sin nx$ au coeficienţii cel puţin de tipul $O\left(\frac{1}{n^{k+1}}\right)$, deci sînt rapid convergente.

<u>Exemplul 1</u>: Să evaluăm cu o eroare $<\varepsilon=10^{-3}$ suma seriei

$$S(x) = \sum_{n=1}^{\infty}\frac{n}{n^2+1}\sin nx .$$

Cum $b_n = \frac{n}{n^2+1}$ seria este lent convergentă. Vom încerca să îmbunătăţim convergenţa acestei serii folosind procedeul indicat mai sus. Pentru aceasta scriem

$$b_n = \frac{n}{n^2+1} = \frac{1}{n}\,\frac{1}{1+\frac{1}{n^2}} = \frac{1}{n}\left(1-\frac{1}{n^2}+\frac{1}{n^4}-\dots\right) = \frac{1}{n}-\frac{1}{n^3}+\tau_n$$

unde $\tau_n = \frac{1}{n^3(n^2+1)}$. Rezultă:

$$\sum_{n=1}^{\infty}\frac{n}{n^2+1}\sin nx = \sum_{n=1}^{\infty}\frac{\sin nx}{n} - \sum_{n=1}^{\infty}\frac{\sin nx}{n^3} + \sum_{n=1}^{\infty}\frac{\sin nx}{n^3(n^2+1)}$$

adică

$$S(x) = \sigma_1(x) + \sigma_2(x) + \sum_{n=1}^{\infty}\tau_n\cdot\sin nx .$$

Fie N - numărul de termeni necesari pentru ca restul R_N să verifice inegalitatea

$$|R_N| = \left|\sum_{n=N+1}^{\infty}\tau_n \sin nx\right| < 10^{-3} .$$

Folosind inegalităţile

$$\left|\sum_{n=N+1}^{\infty}\tau_n\sin nx\right| \le \sum_{n=N+1}^{\infty}\frac{1}{n^5} < \int_N^{\infty}\frac{dx}{x^5} = \frac{1}{4N^4}$$

şi impunînd condiţia $\frac{1}{4N^4}<10^{-3}$ rezultă că este suficient să luăm N = 5 pentru a avea aproximarea cu precizia dorită. Deci

$$S(x) = \frac{\pi-x}{2} - \frac{2\pi^2x-3\pi x^2+x^3}{12} + \sum_{n=1}^{5}\frac{\sin nx}{n^3(n^2+1)} \pm 0{,}001 \qquad (0<x<\pi).$$

Procedura se generalizează pentru serii pentru care a_n sau b_n sînt funcţii raţionale în n, cu gradul numărătorului mai mic cel puţin cu o unitate decît gradul numitorului.

§6. Exerciţii

1. Să se evalueze cu o eroare $< \varepsilon = 10^{-3}$ sumele seriilor

$$\sum_{n=1}^{\infty} \frac{1}{n^3+n+1} \qquad \sum_{n=1}^{\infty} \frac{n}{n^4+1}$$

utilizînd de fiecare dată

a) cele două proceduri prezentate în §1,

b) procedeul lui Kummer .

Determinaţi apoi numărul de termeni necesari şi numărul de zecimale necesare pentru evaluarea aceloraşi sume, prin aceleaşi procedeie, cu eroarea $< \varepsilon = 10^{-4}$ şi respectiv $< \varepsilon = 10^{-5}$.

2. Să se evalueze cu o eroare $< 10^{-3}$ sumele seriilor

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1} n}{n^2+1} \qquad \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} .$$

3. Folosind metoda Euler-Abel calculaţi cu o eroare $< 10^{-3}$ suma seriei

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)^2} x^n$$

pentru $x = -1$.

4. Determinaţi suma seriei $\sum_{n=0}^{\infty} (2n^2-3n+1)x^n$.

5. Evaluaţi cu o eroare $< 10^{-3}$ sumele seriilor:

a) $\sum_{n=1}^{\infty} \frac{n^2+1}{n^3+n^2+1} \sin nx$ b) $\sum_{n=0}^{\infty} \frac{n}{n^2+n+1} \cos nx$

c) $\sum_{n=2}^{\infty} \frac{n^3}{n^4-1} \sin nx$ d) $\sum_{n=1}^{\infty} \frac{n^4-n^2+1}{n^2(n^4+1)} \cos nx$

e) $\sum_{n=1}^{\infty} \frac{n^4}{n^5+1} \sin nx$ f) $\sum_{n=1}^{\infty} \frac{1}{n+a} \sin nx$ $(a > 0)$.

6. Folosind procedeul lui Krîlov, îmbunătăţiţi viteza de convergenţă a seriei trigonometrice asociată funcţiei

$$f(x) = \begin{cases} x^2 + 2 & -\pi \le x < 0 \\ \dfrac{\pi - x}{\pi} & 0 \le x \le \pi \end{cases} .$$

CAPITOLUL III

REZOLVAREA NUMERICA A SISTEMELOR DE ECUATII

A.REZOLVAREA NUMERICA A SISTEMELOR DE ECUATII ALGEBRICE LINIARE

§1.Generalităţi

Considerăm sistemul de m ecuaţii liniare cu n necunoscute:

$$(1.1)\qquad \begin{cases} a_{11}x_1+a_{12}x_2+\dots+a_{1n}x_n = b_1 \\ a_{21}x_1+a_{22}x_2+\dots+a_{2n}x_n = b_2 \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ a_{m1}x_1+a_{m2}x_2+\dots+a_{mn}x_n = b_m \end{cases}$$

unde $a_{ij},b_i \in K$ (K este R sau C), $i=1,2,\dots,m$, $j=1,2,\dots,n$.
Folosind notaţiile:

$$A = \left[a_{ij}\right]_{mxn}, \qquad b = {}^t\left[b_1\dots b_m\right] \qquad X = {}^t\left[x_1\dots x_n\right]$$

sistemul (1.1) se poate scrie sub forma matriceală:

$$(1.2)\qquad A\cdot X = b\ ,$$

Reamintim că sistemul ordonat $(\alpha_1,\alpha_2,\dots,\alpha_n)\in K^n$ se numeşte soluţie a sistemului (1.2) dacă şi numai dacă $A\cdot\alpha = b$, unde $\alpha = {}^t[\alpha_1\ \alpha_2\ \dots\ \alpha_n]$.

Problema de bază a teoriei sistemelor de ecuaţii liniare este determinarea soluţiilor acestui sistem.

Se ştie că problema existenţei soluţiei sistemului (1.1) (sau (1.2)) este complet rezolvată prin teoremele Rouché-Frobenius şi Kronecker-Capelli.Rezolvarea unui sistem compatibil, în cazul particular $m=n=$ rang A (sistem Cramer) se face teoretic cu ajutorul formulelor lui Cramer; rezolvarea sistemelor compatibile cu $m\neq n$ se poate reduce la rezolvarea unui sistem Cramer prin trecerea la un sistem echivalent redus, avînd numărul de ecuaţii egal cu rang A.Din păcate, regula lui Cramer rămîne doar un foarte util instrument teoretic, deoarece calculul valorilor

determinanţilor cu ajutorul calculatorului este o problemă dificilă şi extrem de costisitoare.

Metodele numerice de determinare a soluţiilor sistemului (1.1) sînt de două tipuri:

a) metode directe (exacte)

b) metode iterative.

<u>Metodele directe</u> constau în reducerea sistemului (1.1) la un sistem echivalent rezolvabil direct prin metode elementare. Astfel de metode sînt: regula lui Cramer, metoda eliminării a lui Gauss, metoda rădăcinii pătrate, etc.. Caracteristic pentru metodele directe este faptul că se obţine soluţia exactă efectuînd un număr finit de operaţii, în ipoteza absenţei erorilor de rotunjire.

<u>Metodele iterative</u> constau în construcţia unui şir $\{x^{(k)}\} \subset K^n$ convergent la soluţia exactă a sistemului. Acurateţea soluţiei depinde în acest caz de numărul de iteraţii realizate.

Inainte de a trece la prezentarea metodelor numerice de rezolvare a sistemelor algebrice liniare compatibile, de n ecuaţii cu n necunoscute, remarcăm că următoarele trei probleme sînt strîns legate între ele:

1°. rezolvarea sistemului $A \cdot X = b$

2°. evaluarea lui det A

3°. evaluarea matricei A^{-1}.

Deşi ne vom ocupa în special de rezolvarea primei probleme, vom face scurte referiri şi la problemele 2° şi 3°.

§2. <u>Rezolvarea sistemelor liniare de formă specială</u>

Forma specială a unui sistem liniar corespunde unei forme speciale a matricei A a sistemului: a) matrice diagonală, b) matrice triunghiulară, c) matrice bidiagonală, d) matrice tridiagonală, e) matrice ortogonală, etc..

a) Sisteme cu matrice diagonală. Dacă matricea sistemului este $A = \operatorname{diag}(a_{11}, \dots, a_{nn})$ este nesingulară atunci:

- soluţia sistemului se obţine efectuînd n împărţiri
- determinantul matricei A se obţine efectuînd n-1 înmulţiri
- inversa A^{-1} se obţine făcînd n împărţiri.

b) Sisteme cu matrice triunghiulară. Să presupunem că matricea $A = [a_{ij}]$ a sistemului este nesingulară şi că $a_{ij} = 0$ pentru $i > j$ (adică A este matrice triunghiulară superior). In acest caz soluţia sistemului se poate obţine prin substituţie inversă şi anume:

$$
(2.2)\qquad \begin{cases} x_n = b_n / a_{nn} \\ x_{n-1} = (b_{n-1} - a_{n-1,n} x_n)/a_{n-1,n-1} \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ x_1 = (b_1 - \sum_{j=2}^{n} a_{1j} x_j)/a_{11} . \end{cases}
$$

In acest caz $\det A = \prod_{i=1}^{n} a_{ii}$, iar inversa este tot o matrice triunghiulară superior (vezi teorema 3.1, §3). Pentru rezolvarea sistemului sînt necesare cel mult $n^2 + n$ operaţii; pentru $\det A$ sînt necesare n-1 înmulţiri iar pentru A^{-1} sînt necesare $\leq n^3/6$ înmulţiri.

Cazul sistemelor cu matrice triunghiulară inferior se tratează în mod asemănător. Rezolvarea se face prin substituţie directă (adică se determină întîi x_1, apoi x_2, ş.a.m.d.).

c) Sisteme cu matrice bidiagonală. Pentru a fixa ideile, considerăm că matricea este bidiagonală superior. Acesta este un caz particular al cazului b) şi anume cînd $a_{ij} = 0$ pentru $i > j$ şi pentru $j > i+1$. Se rezolvă folosind substituţia inversă

$$
(2.3)\qquad \begin{cases} x_n = b_n / a_{nn} \\ x_{n-1} = (b_{n-1} - a_{n-1,n} \cdot x_n)/a_{n-1,n-1} \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ x_1 = (b_1 - a_{12} x_2)/a_{11} . \end{cases}
$$

Numărul de operaţii necesare pentru rezolvarea sistemului este $\leq 3n-2$, det A se obţine efectuînd n-1 înmulţiri, iar A^{-1} se obţine efectuînd $\leq n^3/6$ înmulţiri. Este de remarcat faptul că inver-

sa unei matrice bidiagonale superior este o matrice triunghiula-ră superior (în general). De exemplu, inversa matrice $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix}$ este matricea $\begin{bmatrix} 1 & -1 & 5 \\ 0 & 1 & -5 \\ 0 & 0 & 1 \end{bmatrix}$.

d) <u>Sisteme cu matrice tridiagonală</u>. In acest caz sistemul are forma

$$(2.4)\qquad \begin{cases} a_1x_1 + b_1x_2 & = d_1 \\ c_2x_1 + a_2x_2 + b_2x_3 & = d_2 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots & \\ c_{n-1}x_{n-2} + a_{n-1}x_{n-1} + b_{n-1}x_n & = d_{n-1} \\ c_n x_{n-1} + a_n x_n & = d_n \end{cases} .$$

Pentru uşurarea prezentării am notat coeficienţii necunoscute-lor cu cîte un singur indice. Pentru rezolvarea unui astfel de sistem transformăm mai întîi sistemul într-un sistem cu matrice bidiagonală echivalent pe care îl rezolvăm apoi prin substituţie inversă. Vom determina x_1 cu ajutorul lui x_2 din prima ecuaţie. Inlocuim această valoare a lui x_1 în a doua ecuaţie şi putem de-termina x_2 cu ajutorul lui x_3. Procedînd în acest fel cu primele n-1 ecuaţii obţinem relaţii de forma

$$(2,5)\qquad x_r = f_r\, x_{r+1} + g_r , \qquad (r=1,2,..,n-1) .$$

Aici f_r şi g_r sînt încă nedeterminate. Putem acum înlocui în ulti-ma ecuaţie şi determinăm x_n. Apoi, prin substituţie inversă, deter-minăm succesiv $x_{n-1}, x_{n-2}, \dots, x_1$.

Să descriem mai exact acest proces. Există cîteva cazuri de-generate care trebuiesc analizate separat, dar care nu vor fi ana-lizate aici. De exemplu, dacă $a_1=0$, obţinem x_2 şi sistemul se rezol-vă uşor; de asemenea, dacă sistemul este simetric şi un anumit b este zero, atunci sistemul se desface în două subsisteme necupla-te. Ne ocupăm numai de situaţia generală. Din ecuaţia de rang r-1 obţinem $x_{r-1} = f_{r-1}\, x_r + g_{r-1}$ şi, înlocuind în ecuaţia de rang r, obţinem:

$$c_r(f_{r-1}\cdot x_r + g_{r-1}) + a_r\cdot x_r + b_r\cdot x_{r+1} = d_r$$

adică

(2.6) $$x_r = \frac{-b_r}{c_r f_{r-1}+a_r}\, x_{r+1} + \frac{d_r - c_r g_{r-1}}{c_r f_{r-1}+a_r} .$$

Identificînd cu (2,5) obţinem

(2.7) $$f_r = -\frac{b_r}{c_r f_{r-1}+a_r} \qquad g_r = \frac{d_r - c_r g_{r-1}}{c_r f_{r-1}+a_r}$$

($r=2,3,\dots,n-1$). Observăm că $c_1=0$, $b_n=0$ şi că nu avem nevoie să definim f_o, g_o deoarece din prima ecuaţie obţinem

(2.8) $$x_1 = -\frac{b_1}{a_1}\, x_2 + \frac{d_1}{a_1} .$$

Din (2.7) calculăm $f_2,\dots,f_n$ şi $g_2,\dots,g_n$ ţinînd cont că $f_1=-\frac{b_1}{a_1}$, $g_1 = \frac{d_1}{a_1}$. Ultima relaţie dă $f_n=0$ şi $x_n=g_n$, după care obţinem x_{n-1} $x_{n-2},\dots,x_1$ prin substituţie inversă.

Determinantul Δ_n al unei matrice tridiagonale se evaluează uşor printr-o recurenţă pentru trei termeni consecutivi. Dacă punem $\Delta_o=1$, găsim $\Delta_1=a_1$, $\Delta_2= a_1a_2-b_1c_2= a_2\Delta_1-b_1c_2\Delta_o$ şi, în general, dezvoltînd după minorii ultimei linii, obţinem:

(2.9) $$\Delta_r = a_r\Delta_{r-1} - b_{r-1}\, c_r\Delta_{r-2} .$$

Atunci Δ_n se poate calcula cu $\leq 3(n-1)$ înmulţiri şi n-1 adunări, adică cu $\leq 4(n-1)$ operaţii.

In legătură cu inversa unei matrice tridiagonale facem observaţia că poate avea toate elementele nenule. De exemplu, inversa matricei $\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}$ este matricea $\begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$,

Observăm că ideia de rezolvare a problemei în cazul matricei tridiagonale se poate adapta la matrice 5-diagonale şi, mai general, la matrice (2m+1)-diagonale (cu $2m+1<n$). Aceste matrice se mai numesc şi matrice bandă şi sînt caracterizate de condiţia $a_{ij} = 0$ dacă $|i-j|> m$, $2m+1 < n$.

e) <u>Sisteme cu matrice ortogonală</u>. Dacă matricea A a sistemului este ortogonală, adică $A^{-1} = {}^tA$, atunci sistemul $A\cdot X = b$ are soluţia $X = {}^tA.b$.

<u>Observaţia 2.1</u>: Deoarece rezolvarea sistemelor de formă specială se face printr-un număr relativ mic de operaţii aritme-

tice,apare problema reducerii sistemelor (1.1) la sisteme de formă specială.Vom arăta că în condiţii destul de fireşti este posibil de redus sistemele liniare compatibile la sisteme de forme particulare indicate mai sus.In o mare parte din cazuri se realizează factorizarea de diverse tipuri a matricei sistemului.Tridiagonalizarea se realizează şi printr-un procedeu de interes deosebit,diferit de tehnicile obişnuite de factorizare,ce va fi prezentat separat.

§3.Factorizarea matricelor.Inversa generalizată a unei matrice

Demonstraţiile teoremelor de factorizare ale matricelor care vor fi prezentate aici vor folosi inducţia completă şi,implicit,operaţii cu matrice împărţite în blocuri.

Impărţirea unei matrice în blocuri (submatrice) se realizează folosind linii punctate verticale şi orizontale ca în exemplele de mai jos:

$$A = \left[\begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{array}\right], B = \left[\begin{array}{cc|cc|c} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} \\ b_{31} & b_{32} & b_{33} & b_{34} & b_{35} \\ \hline b_{41} & b_{42} & b_{43} & b_{44} & b_{45} \\ b_{51} & b_{52} & b_{53} & b_{54} & b_{55} \end{array}\right]$$

Matricele A şi B pot fi privite ca matrice ale căror elemente sînt tot matrice.Numerotarea blocurilor unei matrice împărţite în blocuri se realizează în acelaşi mod ca numerotarea elementelor unei matrice obişnuite.De exemplu,pentru matricele A şi B de mai sus avem:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{bmatrix}$$

unde

$$A_{11} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \qquad A_{12} = \begin{bmatrix} a_{14} & a_{15} \\ a_{24} & a_{25} \end{bmatrix}$$

$$A_{21} = \begin{bmatrix} a_{31} & a_{32} & a_{33} \end{bmatrix} \qquad A_{22} = \begin{bmatrix} a_{34} & a_{35} \end{bmatrix}$$

$$B_{11} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} \qquad B_{12} = \begin{bmatrix} b_{13} & b_{14} \\ b_{23} & b_{24} \\ b_{33} & b_{34} \end{bmatrix} \qquad B_{13} = \begin{bmatrix} b_{15} \\ b_{25} \\ b_{35} \end{bmatrix}$$

$$B_{21} = \begin{bmatrix} b_{41} & b_{42} \\ b_{51} & b_{52} \end{bmatrix} \qquad B_{22} = \begin{bmatrix} b_{43} & b_{44} \\ b_{53} & b_{54} \end{bmatrix} \qquad B_{23} = \begin{bmatrix} b_{45} \\ b_{55} \end{bmatrix} .$$

Ca matrice cu blocuri,cele două matrice au dimensiunile 2 x 2 şi respectiv 2 x 3.Prin calcul direct se verifică egalitatea:

$$A \cdot B = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}$$

adică,dacă matricele cu blocuri sînt înlănţuite şi,la rîndul lor,blocurile coloanei i a primei matrice sînt înlănţuite cu blocurile liniei i a matricei a 2-a (astfel încît şi adunările de matrice care apar să fie permise) atunci produsul matricelor cu blocuri se realizează formal după aceleaşi reguli ca produsul a două matrice înlănţuite obişnuite.In acest caz vom spune că împărţirea în blocuri a celor două matrice este compatibilă cu înmulţirea matricelor cu blocuri.

De asemenea,suma a două matrice cu blocuri de acelaşi tip (adică matrice care au aceleaşi dimensiuni ca matrice cu blocuri,iar blocurile de pe locuri corespunzătoare au aceleaşi dimensiuni) se realizează formal după aceeaşi regulă ca suma a două matrice obişnuite de aceleaşi dimensiuni.De exemplu:

$$A + C = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11}+C_{11} & A_{12}+C_{12} \\ A_{21}+C_{21} & A_{22}+C_{22} \end{bmatrix}$$

dacă blocurile A_{ij} şi C_{ij} $(i,j=1,2)$ au aceleaşi dimensiuni.

In cele ce urmează ne sînt utile următoarele notaţii:

$I^{(n)}$ = mulţimea matricelor pătratice,de ordinul n,nesingulare, triunghiulare inferior

$I_1^{(n)}$ = mulţimea matricelor pătratice de ordinul n,triunghiulare inferior,care au toate elementele diagonalei principale egale cu 1

$D^{(n)}$ = mulţimea matricelor diagonale de ordinul n,nesingulare

$S^{(n)}$ = mulţimea matricelor pătratice de ordinul n,nesingulare, triunghiulare superior

$S_1^{(n)}$ = mulţimea matricelor pătratice de ordinul n, triunghiulare superior, care au toate elementele diagonalei principale egale cu 1.

Are loc

<u>Teorema 3.1</u>: $I^{(n)}, I_1^{(n)}, D^{(n)}, S^{(n)}$ <u>şi</u> $S_1^{(n)}$ <u>sînt subgrupuri ale grupului liniar general</u> $GL_n(K)$.

<u>Demonstraţie</u>: Vom justifica numai că $I^{(n)}$ este subgrup al lui $GL_n(K)$. Similar se poate justifica afirmaţia cu privire la $I_1^{(n)}, D^{(n)}, S^{(n)}$ şi $S_1^{(n)}$.

Fie $A = [a_{ij}] \in I^{(n)}$ şi $B = [b_{ij}] \in I^{(n)}$. Atunci $a_{ij} = b_{ij} = 0$ pentru $i < j$. Dar atunci, pentru $i < j$, au loc egalităţile:

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=1}^{i} a_{ik}b_{kj} + \sum_{k=i+1}^{n} a_{ik}b_{kj} = 0$$

(deoarece în prima sumă $k \leq i < j$, deci $b_{kj} = 0$, iar în a doua sumă $i < k$ adică $a_{ik} = 0$). In plus

$$c_{ii} = \sum_{k=1}^{i-1} a_{ik}b_{ki} + a_{ii}b_{ii} + \sum_{k=i+1}^{n} a_{ik}b_{ki} = a_{ii}b_{ii} \quad (i=1,2,\dots,n)$$

Deci $A \cdot B \in I^{(n)}$. Vom dovedi, prin inducţie completă după ordinul n al matricei, că inversa unei matrice triunghiulare inferior este tot o matrice triunghiulară inferior. Afirmaţia este evidentă pentru orice matrice triunghiulară inferior de ordinul $n = 1$ (nesingulară). Presupunem că $I^{(n-1)}$ este subgrup al grupului $GL_{n-1}(K)$. Dacă $A = [a_{ij}] \in I^{(n)}$, o putem scrie sub forma de matrice cu blocuri

$$A = \begin{bmatrix} a_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$$

unde $A_{22} \in I^{(n-1)}$. Căutăm inversa $B = A^{-1}$ tot sub formă de matrice cu blocuri de tipul

$$B = \begin{bmatrix} b_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} .$$

Atunci condiţia

$$A \cdot B = I_n = \begin{bmatrix} 1 & 0 \\ 0 & I_{n-1} \end{bmatrix}$$

(I_k notează matricea unitate de ordinul k) ne conduce la egalităţile:

$$a_{11}\cdot b_{11} = 1 \qquad A_{21}\cdot b_{11} + A_{22}B_{21} = 0$$
$$a_{11}\cdot B_{12} = 0 \qquad A_{21}\cdot B_{12} + A_{22}\cdot B_{22} = I_{n-1}.$$

Dar $A \in I^{(n)}$ implică $a_{11} \neq 0$ şi prin urmare $B_{12} = 0$. Ultima relaţie devine $A_{22}\cdot B_{22} = I_{n-1}$. Conform ipotezei inductive $B_{22} \in I^{(n-1)}$ şi prin urmare $B = A^{-1} \in I^{(n)}$. *

<u>Definiţia 3.1</u>: <u>Fie</u> $A = [a_{ij}] \in \mathcal{M}_n(K)$. <u>Minorii</u> $\det\begin{pmatrix}1 & 2 & ..\ell\\ 1 & 2 & ..\ell\end{pmatrix}_A$, $1 \leq \ell \leq n$ se <u>numesc minorii diagonali principali ai matricei</u> A.*

Reamintim că $\mathcal{M}_n(K)$ notează mulţimea matricelor de tip $n \times n$ cu elemente din cîmpul K şi că $\det\begin{pmatrix}i_1 & i_2..i_k\\ j_1 & j_2..j_k\end{pmatrix}$ notează minorul obţinut la intersecţia liniilor $i_1, i_2, .., i_k$ cu coloanele $j_1, j_2, .., j_k$.

<u>Definiţia 3.2</u>: <u>Se spune că matricea pătratică</u> $A \in \mathcal{M}_n(K)$ <u>are proprietatea</u> K_n <u>dacă toţi minorii diagonali principali ai săi sînt</u> $\neq 0$.

Are loc:

<u>Teorema 3.2</u>: <u>Orice matrice pătratică</u> $A \in \mathcal{M}_n(K)$ <u>care poate fi pusă sub forma</u> $A = B\cdot D\cdot C$ <u>cu</u> $B \in I_1^{(n)}$, $D \in D^{(n)}$ <u>şi</u> $C \in S_1^{(n)}$ <u>are proprietatea</u> K_n.*

<u>Demonstraţie</u>: Afirmaţia este evidentă pentru orice matrice de ordinul I (într-adevăr, $[\alpha] = [1]\cdot[\alpha]\cdot[1]$). O presupunem adevărată pentru orice matrice pătratică de ordinul n-1 care satisface ipoteza teoremei. Pentru demonstraţie, scriem matricele A, B, C, $D \in \mathcal{M}_n(K)$ sub forma:

$$A = \begin{bmatrix} A_{11} & A_{12}\\ A_{21} & a_{nn}\end{bmatrix},\ B = \begin{bmatrix} B_{11} & 0\\ B_{21} & 1\end{bmatrix},\quad C = \begin{bmatrix} C_{11} & C_{12}\\ 0 & 1\end{bmatrix},\ D = \begin{bmatrix} D_1 & 0\\ 0 & d_n\end{bmatrix}$$

unde $B_{11} \in I_1^{(n-1)}$, $C_{11} \in S_1^{(n-1)}$, $D_1 \in D^{(n-1)}$. Atunci egalitatea $A = B.D\cdot C$ devine:

$$A_{11} = B_{11}\cdot D_1\cdot C_{11} \qquad A_{12} = B_{11}\cdot D_1\cdot C_{12}$$
$$A_{21} = B_{21}\cdot D_1\cdot C_{11} \qquad a_{nn} = B_{21}\cdot D_1\cdot C_{12} + d_n.$$

Conform ipotezei inductive, minorii diagonali principali ai matricei A_{11} sînt nenuli. Pentru a dovedi că matricea A satisface

proprietatea K_n este suficient să dovedim că $\det A \neq 0$. Dar $\det A = \det B \cdot \det D \cdot \det C \neq 0$ (deoarece $D \in D^{(n)}$, iar $\det B = \det C = 1$).*

Vom dovedi acum reciproca teoremei 3.2 şi anume:

<u>Teorema 3.3</u>: <u>Orice matrice</u> $A \in \mathcal{M}_n(K)$ <u>care are proprietatea K_n poate fi scrisă în mod unic sub forma</u> $A = B \cdot D \cdot C$ <u>cu</u> $B \in I_1^{(n)}$, $D \in D^{(n)}$ <u>şi</u> $C \in S_1^{(n)}$.

<u>Demonstraţie</u>: Folosim din nou inducţia completă după n. Afirmaţia teoremei are loc pentru orice matrice de ordinul 1 şi presupunem că are loc şi pentru orice matrice de ordinul n-1 care satisface proprietatea K_{n-1}. Fie $A = [a_{ij}] \in \mathcal{M}_n(K)$ o matrice care satisface proprietatea K_n. O putem scrie sub formă de matrice cu blocuri:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{nn} \end{bmatrix} .$$

Căutăm matricele B, D şi C care satisfac condiţiile din enunţul teoremei tot sub formă de matrice cu blocuri:

$$B = \begin{bmatrix} B_{11} & 0 \\ B_{21} & 1 \end{bmatrix} \quad D = \begin{bmatrix} D_1 & 0 \\ 0 & d_n \end{bmatrix} \quad C = \begin{bmatrix} C_{11} & C_{12} \\ 0 & 1 \end{bmatrix}$$

unde $B_{11} \in I_1^{(n-1)}$, $D_1 \in D^{(n-1)}$ şi $C_{11} \in S_1^{(n-1)}$. Impunînd condiţia $A = B \cdot D \cdot C$ obţinem:

$$A_{11} = B_{11} \cdot D_1 \cdot C_{11} \qquad A_{12} = B_{11} \cdot D_1 \cdot C_{12}$$
$$A_{21} = B_{21} \cdot D_1 \cdot C_{11} \qquad a_{nn} = B_{21} \cdot D_1 \cdot C_{12} + d_n .$$

Deoarece matricea A_{11} satisface proprietatea K_{n-1}, conform ipotezei inductive vor exista matricele $B_{11} \in I_1^{(n-1)}$, $D_1 \in D^{(n-1)}$ şi $C_{11} \in S_1^{(n-1)}$ astfel încît $A_{11} = B_{11} \cdot D_1 \cdot C_{11}$. Deoarece $\det(B_{11} D_1) = \det(D_1 C_{11}) = \det D_1 \neq 0$, rezultă că există matricele $(B_{11} D_1)^{-1}$ şi $(D_1 C_{11})^{-1}$ şi putem determina matricele

$$C_{12} = (B_{11} D_1)^{-1} A_{12} \qquad B_{21} = A_{21} (D_1 C_{11})^{-1}$$
$$d_n = a_{nn} - A_{21} A_{11}^{-1} A_{12} .$$

Prin urmare avem asigurată existenţa matricelor B, D şi C care să satisfacă condiţiile teoremei. Să dovedim unicitatea acestor matrice. Presupunem că ar mai exista nişte matrice $B' \in I_1^{(n)}$, $D' \in D^{(n)}$ şi $C' \in S_1^{(n)}$ astfel încît $A = B' \cdot D' \cdot C'$. Egalitatea $B \cdot D \cdot C = B' \cdot D' \cdot C'$ ne spune că $(B'D')^{-1} \cdot (BD) = C' \cdot C^{-1} \in I_1^{(n)} \cap S_1^{(n)} = \{I_n\}$

şi prin urmare $C = C'$ şi $BD = B'D'$. Ultima egalitate se poate scrie sub forma $B'^{-1}B = D'D^{-1} \in I_1^{(n)} \cap D^{(n)} = \{I_n\}$, adică $B = B'$ şi $D = D'$.*

<u>Definiţia 3.3</u>: <u>Dacă pentru matricea</u> $A \in \mathcal{M}_n(K)$ <u>care satisface proprietatea</u> K_n <u>s-au determinat matricele</u> $B \in I_1^{(n)}$, $D \in D^{(n)}$ <u>şi</u> $C \in S_1^{(n)}$ <u>cu proprietatea</u> $A = B\cdot D\cdot C$, <u>spunem că am realizat</u> <u>L.D.U-factorizarea</u> (<u>sau I.D.S.-factorizarea</u>) <u>matricei A</u>.*

<u>Exemplul 3.1</u>: Să determinăm L.D.U.-factorizarea matricei

$$A = \begin{bmatrix} 2 & 1 & -1 & 1 \\ 4 & 0 & -2 & 1 \\ -2 & -3 & 2 & -4 \\ 0 & -2 & -1 & 3 \end{bmatrix}.$$

Matricea A are minorii diagonali principali $\Delta_1 = 2$, $\Delta_2 = -4$, $\Delta_3 = -4$, $\Delta_4 = -8$, deci satisface proprietatea K_4. Căutăm matricele

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ b_{21} & 1 & 0 & 0 \\ b_{31} & b_{32} & 1 & 0 \\ b_{41} & b_{42} & b_{43} & 1 \end{bmatrix}, D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix}, C = \begin{bmatrix} 1 & c_{12} & c_{13} & c_{14} \\ 0 & 1 & c_{23} & c_{24} \\ 0 & 0 & 1 & c_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

care satisfac egalitatea $A = B\cdot D\cdot C$. Sîntem conduşi la următoarele relaţii:

$$\begin{array}{lll} d_1 = 2 & d_1c_{12} = 1 & d_1c_{13} = -1 \\ b_{21}d_1 = 4 & d_1c_{12}b_{21} + d_2 = 0 & b_{21}d_1c_{13} + d_2c_{23} = -2 \\ b_{31}d_1 = -2 & b_{31}d_1c_{12} + b_{32}d_2 = -3 & b_{31}d_1c_{13} + b_{32}d_2c_{23} + d_3 = 2 \\ b_{41}d_1 = 0 & b_{41}d_1c_{12} + b_{42}d_2 = -2 & b_{41}d_1c_{13} + b_{42}d_2c_{23} + b_{43}d_3 = -1 \end{array}$$

$$\begin{array}{l} d_1c_{14} = 1 \\ b_{21}d_1c_{14} + d_2c_{24} = 1 \\ b_{31}d_1c_{14} + b_{32}d_2c_{24} + d_3c_{34} = -4 \\ b_{41}d_1c_{14} + b_{42}d_2c_{24} + b_{43}d_3c_{34} + d_4 = 3 \end{array}$$

care ne conduc la matricele

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 \end{bmatrix} \quad D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad C = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

<u>Consecinţa 3.1</u>: <u>Orice matrice</u> $A \in S^{(n)}$ <u>se scrie în mod unic sub forma</u> $A = D\cdot C$ <u>cu</u> $D \in D^{(n)}$ <u>şi</u> $C \in S_1^{(n)}$.*

<u>Consecinţa 3.2</u>: <u>Orice matrice</u> $A \in I^{(n)}$ <u>se scrie în mod unic sub forma</u> $A = B\cdot D$ <u>cu</u> $B \in I_1^{(n)}$ <u>şi</u> $D \in D^{(n)}$.*

Observaţia 3.1:In practică, în locul L.D.U.-factorizării se foloseşte L.U.-factorizarea (a se vedea observaţia 3.2). In acest caz, matricea $A = [a_{ij}] \in \mathcal{M}_n(K)$ se scrie sub forma $A = B \cdot C$ cu $B = [b_{ij}] \in I^{(n)}$ şi $C = [c_{ij}] \in S^{(n)}$. Inmulţind matricele B şi C şi egalînd rezultatul cu A, obţinem relaţiile: $\sum_{k=1}^{\min(i,j)} b_{ik}c_{kj} = a_{ij}$ sau

$$\begin{array}{lll}
b_{11}c_{11}=a_{11} & b_{21}c_{11}=a_{21} & \ldots\ldots b_{r1}c_{11}=a_{r1} \\
b_{11}c_{12}=a_{12} & b_{21}c_{12}+b_{22}c_{22}=a_{22} & b_{r1}c_{12}+b_{r2}c_{22}=a_{r2} \\
\ldots\ldots & \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\
b_{11}c_{1n}=a_{1n} & b_{21}c_{1n}+b_{22}c_{2n}=a_{2n} & b_{r1}c_{1r}+\ldots+b_{rr}c_{rr}=a_{rr}
\end{array}$$

cu $r \leq n$. Acest sistem se rezolvă unic dacă se dau valorile b_{rr}. De obicei se ia $b_{rr}=1$. Dacă s-a realizat L.U.-factorizarea se poate trece uşor la L.D.U.-factorizare scriind $B=B'D_1$, $C=D_2C'$ cu $D_1 = \operatorname{diag}(b_{11},\ldots,b_{nn})$, $D_2 = \operatorname{diag}(c_{11},\ldots,c_{nn})$, $B' \in I_1^{(n)}$ şi $C' \in S_1^{(n)}$; în acest caz $A = B'DC'$ cu $D = D_1D_2$.*

Teoremele 3.1 şi 3.2 arată că clasa matricelor care admit L.D.U.-factorizare este caracterizată de proprietatea K_n. Teorema care urmează ne arată cît de largă este această clasă:

Teorema 3.4: Dacă matricea pătratică $A \in \mathcal{M}_n(K)$ este nesingulară, atunci, schimbînd eventual ordinea liniilor (sau coloanelor) şi, se ajunge la o matrice care are proprietatea K_n.

Demonstraţie: Dacă $\det A \neq 0$ şi A nu satisface proprietatea K_n, considerăm numărul j astfel încît

$$\det\begin{pmatrix}1 & 2 & \ldots & i\\ 1 & 2 & \ldots & i\end{pmatrix}_A \neq 0 \qquad 1 \leq i \leq j-1$$

şi

$$\det\begin{pmatrix}1 & 2 & \ldots & j\\ 1 & 2 & \ldots & j\end{pmatrix}_A = 0 .$$

Va exista atunci un număr $k > j$ astfel încît $\det\begin{pmatrix}1 & 2 \ldots j-1 & k\\ 1 & 2 \ldots j-1 & j\end{pmatrix}_A \neq 0$ deoarece, în caz contrar, coloanele $\{C_1^A, C_2^A, \ldots, C_j^A\}$ ale matricei A ar fi liniar dependente, fapt care ar contrazice ipoteza $\det A \neq 0$. Schimbînd între ele liniile k şi j se obţine o matrice B pentru care $\det\begin{pmatrix}1 & 2 \ldots j\\ 1 & 2 \ldots j\end{pmatrix}_B = \det\begin{pmatrix}1 & 2 \ldots j-1 & k\\ 1 & 2 \ldots j-1 & j\end{pmatrix}_A \neq 0$. După cel mult $n-j+1$ astfel de schimbări de linii se ajunge la o matrice care are proprietatea K_n.*

Exemplul 3.2: Considerăm matricea nesingulară

$$A = \begin{bmatrix} 0 & 1 & -2 & 3 \\ 2 & 1 & 0 & 1 \\ 2 & 2 & -2 & 3 \\ 1 & 0 & -1 & 2 \end{bmatrix} \qquad (\det A = 4).$$

Deoarece $\det \begin{pmatrix} 1 \\ 1 \end{pmatrix}_A = 0$ şi $\det \begin{pmatrix} 2 \\ 1 \end{pmatrix}_A \neq 0$ schimbăm între ele liniile 1 şi 2 şi obţinem matricea

$$A_1 = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 0 & 1 & -2 & 3 \\ 2 & 2 & -2 & 3 \\ 1 & 0 & -1 & 2 \end{bmatrix}$$

Pentru A_1 primii doi minori diagonali principali sînt $\neq 0$, dar $\det \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}_{A_1} = 0$. Cum $\det \begin{pmatrix} 1 & 2 & 4 \\ 1 & 2 & 3 \end{pmatrix}_{A_1} \neq 0$, schimbînd între ele liniile 3 şi 4 obţinem matricea

$$B = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 0 & 1 & -2 & 3 \\ 1 & 0 & -1 & 2 \\ 2 & 2 & -2 & 3 \end{bmatrix}$$

care are proprietatea K_4.

Teorema 3.5: Orice matrice simetrică $A \in \mathcal{M}_n(K)$ (cu $A = {}^tA$) care are proprietatea K_n poate fi scrisă în mod unic sub forma $A = {}^tC \cdot D \cdot C$ cu $C \in S_1^{(n)}$ şi $D \in D^{(n)}$.

Demonstraţie: Conform teoremei 3.3 putem scrie în mod unic A sub forma $A = B \cdot D \cdot C$ cu $B \in I_1^{(n)}$, $D \in D^{(n)}$ şi $C \in S_1^{(n)}$. Dar egalităţile $A = {}^tA$ şi $D = {}^tD$ ne conduc la

$$A = B \cdot D \cdot C = {}^t(B \cdot D \cdot C) = {}^tC \cdot {}^tD \cdot {}^tB = {}^tC \cdot D \cdot {}^tB.$$

Datorită unicităţii L.D.U.-factorizării matricei A, rezultă că $B = {}^tC$ şi $C = {}^tB$, adică $A = {}^tC \cdot D \cdot C$. *

Exemplul 3.3: Se dă matricea simetrică

$$A = \begin{bmatrix} 2 & 4 & 0 & 2 \\ 4 & 9 & -2 & 6 \\ 0 & -2 & 3 & -3 \\ 2 & 6 & -3 & 3 \end{bmatrix}.$$

Ne propunem să o punem sub forma $A = {}^tC \cdot D \cdot C$. Remarcăm mai întîi că A are proprietatea K_4. Fie

$$C = \begin{bmatrix} 1 & c_{12} & c_{13} & c_{14} \\ 0 & 1 & c_{23} & c_{24} \\ 0 & 0 & 1 & c_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix}$$

Condiţia $A = {}^tC \cdot D \cdot C$ conduce la egalităţile

$$\begin{array}{llll} d_1=2 & d_1c_{12}=4 & d_1c_{13}=0 & d_1c_{14}=2 \\ & c_{12}d_1c_{12}+d_2=9 & c_{12}d_1c_{13}+c_{23}d_2=-2 & c_{12}d_1c_{14}+c_{24}d_2=6 \\ & & c_{13}d_1c_{13}+c_{23}d_2c_{23}=3 & c_{13}d_1c_{14}+c_{23}d_2c_{24}+ \\ & & & +c_{34}d_3=-3 \end{array}$$

$$c_{14}d_1c_{14}+c_{24}d_2c_{24}+c_{34}d_3c_{34}+d_4=3.$$

Rezolvînd acest sistem se obţin matricele:

$$C = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & -2 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}.$$

Teorema 3.6: Orice matrice pătratică simetrică $A \in \mathcal{M}_n(C)$ (C - corpul numerelor complexe) care are proprietatea K_n se poate factoriza sub forma $A = {}^tS \cdot S$ cu $S \in S^{(n)}$.

Demonstraţie: Conform teoremei precedente, matricea A se poate factoriza în mod unic sub forma $A = {}^tC \cdot D \cdot C$ cu $C \in S_1^{(n)}$ şi $D \in D^{(n)}$. Căutăm matricea diagonală $F \in \mathcal{M}_n(C)$ astfel încît $F^2 = D$. Dacă $F = \mathrm{diag}(x_1, x_2, \dots, x_n)$, atunci $F^2 = \mathrm{diag}(x_1^2, x_2^2, \dots, x_n^2)$, iar egalitatea $F^2 = D$ conduce la condiţiile $x_i^2 = d_i$ $(i=1,2,\dots,n)$ adică $x_i = \pm\sqrt{d_i}$. Deoarece nu există un criteriu natural de alegere a semnelor elementelor x_i, rezultă că, corespunzător diverselor alegeri ale acestor semne, vom obţine 2^n matrice diagonale F care satisfac condiţia $F^2 = D$. Fixîndu-ne la o alegere particulară a semnelor (de exemplu să alegem $x_i = \sqrt{d_i}$, $i=1,2,\dots,n$), obţinem o matrice F pentru care $F^2 = {}^tF \cdot F = D$ şi, în plus:

$$A = {}^tC \cdot {}^tF \cdot F \cdot C = {}^t(F \cdot C) \cdot (F \cdot C) = {}^tS \cdot S$$

unde $S = F \cdot C \in S^{(n)}$. *

Definiţia 3.4: Dacă matricea simetrică $A \in \mathcal{M}_n(C)$ a fost pusă sub forma $A = {}^tS \cdot S$ cu $S \in S^{(n)}$, spunem că am realizat o factorizare Cholesky a matricei A. *

Exemplul 3.4: Să determinăm o factorizare Cholesky a matricei simetrice A din exemplul 3.3. In exemplul 3.3 am determinat matricele $C \in S^{(4)}$ şi $D \in D^{(4)}$ astfel încît $A = {}^tC \cdot D \cdot C$. Conform demonstraţiei teoremei 3.6 rămîne să determinăm matri-

cea $F = \mathrm{diag}(x_1,x_2,x_3,x_4)$ din condiţia $F^2 = D$. Cum $D=\mathrm{diag}(2;1;-1;-2)$, condiţia $F^2 = D$ implică:

$$x_1 = \pm\sqrt{2} \quad , \quad x_2 = \pm 1 \quad , \quad x_3 = \pm j \quad , \quad x_4 = \pm j\sqrt{2} \quad (j^2=-1).$$

Să alegem $F = \mathrm{diag}(\sqrt{2}\ ;1;j;j\sqrt{2})$. Obţinem:

$$S = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & j & 0 \\ 0 & 0 & 0 & j\sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & -2 & 2 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 2\sqrt{2} & 0 & \sqrt{2} \\ 0 & 1 & -2 & 2 \\ 0 & 0 & j & -j \\ 0 & 0 & 0 & j\sqrt{2} \end{bmatrix} .$$

Se vede că, deşi matricea are elemente reale, factorizarea Cholesky se realizează cu matrice $S \in S^{(4)}$ cu elemente complexe. Mai exact, orice factorizare Cholesky a matricei A se realizează cu matrice factori cu elemente complexe.*

Apare problema: Să se stabilească în ce caz factorizările Cholesky ale unei matrice simetrice cu elemente reale se realizează cu matrice factori $S \in S^{(n)}$ cu elemente reale.

Răspunsul la această problemă îl dă teorema următoare:

<u>Teorema 3.7</u>: <u>Condiţia necesară şi suficientă ca o matrice pătratică simetrică A, cu elemente reale, să admită o factorizare Cholesky cu matrice reale ca factori este ca toţi minorii diagonali principali ai matricei A să fie pozitivi.</u>

<u>Necesitatea</u>: Presupunem că matricea pătratică simetrică A, cu elemente reale, admite o factorizare Cholesky $A = {}^tS\cdot S$ unde $S \in S^{(n)}$ are elemente reale. Folosind inducţia completă după ordinul n al matricei, vom arăta că toţi minorii diagonali principali sînt pozitivi. Afirmaţia este evident adevărată pentru $n=1$. O presupunem adevărată pentru matrice simetrice, cu elemente reale, de ordinul $n-1$. Fie $A \in \mathcal{M}_n(R)$, o matrice simetrică care admite factorizarea Cholesky $A = {}^tS.S$ cu $S \in S^{(n)}$, S avînd elemente reale. Punînd matricele A şi S sub formă de matrice cu blocuri

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{nn} \end{bmatrix} \qquad S = \begin{bmatrix} S_{11} & S_{12} \\ 0 & s_{nn} \end{bmatrix}$$

şi, ţinînd cont că $A_{21} = {}^tA_{12}$, relaţia $A = {}^tS\cdot S$ ne conduce la egalităţile:

$$A_{11} = {}^tS_{11}\cdot S_{11} \qquad A_{12} = {}^tS_{11}\cdot S_{12}$$
$$a_{nn} = {}^tS_{12}\cdot S_{12} + s_{nn}^2 .$$

Deoarece matricea simetrică A_{11} admite factorizare Cholesky cu factori reali şi are ordinul n-1, conform ipotezei inductive, minorii diagonali principali ai săi sînt pozitivi. Dar aceşti minori sînt primii n-1 minori diagonali principali ai matricei A. Rămîne deci de dovedit că $\det A > 0$. Dar $\det A = \det({}^tS\cdot S) = (\det S)^2 > 0$.

Suficienţa: Presupunem că matricea simetrică $A \in \mathcal{M}_n(R)$ are toţi minorii diagonali principali pozitivi. Conform teoremei 3.5 matricea A se poate factoriza în mod unic sub forma $A = {}^tC\cdot D\cdot C$ cu $D \in D^{(n)}$ şi $C \in S_1^{(n)}$. Pentru a justifica suficienţa condiţiei din teoremă este suficient să arătăm că D are elementele de pe diagonala principală strict pozitive. Folosim din nou inducţia completă după dimensiunea n a matricei A. Pentru matrice de ordinul n=1 afirmaţia este evidentă. O presupunem adevărată pentru matrice simetrice de ordinul n-1 (care au minorii diagonali principali pozitivi). Scriem matricele A, C şi D sub formă de matrice cu blocuri:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & a_{nn} \end{bmatrix} \qquad C = \begin{bmatrix} C_{11} & C_{12} \\ 0 & 1 \end{bmatrix} \qquad D = \begin{bmatrix} D_1 & 0 \\ 0 & d_n \end{bmatrix}$$

(cu $A_{21} = {}^tA_{12}$). Egalitatea $A = {}^tC\cdot D\cdot C$ conduce la egalităţile:

$$A_{11} = {}^tC_{11}\cdot D_1\cdot C_{11} \qquad A_{12} = {}^tC_{11}\cdot D_1\cdot C_{12}$$
$$A_{21} = {}^tA_{12} \qquad a_{nn} = {}^tC_{12}\cdot D_1\cdot C_{12} + d_n .$$

Conform ipotezei inductive rezultă că elementele diagonalei principale a matricei D_1 sînt strict pozitive. Rămîne de dovedit că $d_n > 0$. Dar $\det A = (\det C)^2\cdot \det D = \det D = (\det D_1)\cdot d_n$, adică $d_n = \frac{\det A}{\det D_1} > 0$. *

Exemplul 3.5: Considerăm matricea simetrică $A \in \mathcal{M}_4(R)$:

$$A = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 1 & 5 & -1 & 4 \\ -1 & -1 & 5 & -2 \\ 2 & 4 & -2 & 6 \end{bmatrix} .$$

Minorii diagonali principali ai matricei A sînt:

$\Delta_1 = 1$, $\Delta_2 = 4$, $\Delta_3 = 16$, $\Delta_4 = \det A = 72$.

Deoarece minorii diagonali principali sînt toţi pozitivi,rezultă că matricea A admite o factorizare Cholesky cu factori reali. Pentru a determina o factorizare Cholesky a matricei A,căutăm $S \in S^{(4)}$ de forma:

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ 0 & s_{22} & s_{23} & s_{24} \\ 0 & 0 & s_{33} & s_{34} \\ 0 & 0 & 0 & s_{44} \end{bmatrix}$$

Impunînd condiţia $A = {}^tS \cdot S$ se ajunge la sistemul

$$\begin{array}{llll} s_{11}^2 = 1 & s_{11}s_{12} = 1 & s_{11}s_{13} = -1 & s_{11}s_{14} = 2 \\ & s_{12}^2 + s_{22}^2 = 5 & s_{12}s_{13} + s_{22}s_{23} = -1 & s_{12}s_{14} + s_{22}s_{24} = 4 \\ & & s_{13}^2 + s_{23}^2 + s_{33}^2 = 5 & s_{13}s_{14} + s_{23}s_{24} + s_{33}s_{34} = \\ & & & s_{14}^2 + s_{24}^2 + s_{34}^2 + s_{44}^2 = 6 \end{array}$$

care admite ca soluţie elementele s_{ij} $(i \leq j)$ ale matricei

$$S = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

Dacă am fi impus condiţia ca toate elementele diagonalei principale ale matricei S să fie pozitive,am fi ajuns la factorizarea $A = {}^tS_1 \cdot S_1$ cu

$$S_1 = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.^*$$

Prin teorema 3.7 am determinat clasa amtricelor simetrice,cu elemente reale care admit factorizare Cholesky cu matrice factori cu elemente reale.Demonstraţia teoremei (vezi şi exemplul 3.5) nu asigură unicitatea unei astfel de factorizări (în caz că este posibilă);demonstraţia teoremei 3.6 precizează că în general există 2^n factorizări Cholesky pentru orice matrice simetrică.Dacă impunem condiţia ca elementele diagonalei principale a matricei factor S să fie strict pozitive,vom avea o factorizare Cholesky unică.Deoarece pe acest tip de factorizare se bazează un procedeu de rezolvare numerică (des utilizat) al u-

nui sistem de ecuaţii liniare, vom indica procedeul concret de determinare a factorizării Cholesky.

Fie deci $A = a_{ij} \in \mathcal{M}_n(R)$ cu $a_{ij}=a_{ji}$, avînd minorii diagonali principali pozitivi. Dacă $A = {}^tS\cdot S$, unde $S = [s_{ij}] \in \mathcal{M}_n(R)$ cu $s_{ij}=0$ pentru $i > j$ şi $s_{ii} > 0$ $(i=1,2,..,n)$, atunci:

$$
(3.1)\quad
\begin{cases}
s_{11} = \sqrt{a_{11}} \qquad s_{1i} = a_{1i}/s_{11} & i=2,3,\ldots,n \\
s_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} s_{ki}^2\right)^{\frac{1}{2}} & i=2,3,\ldots,n \\
s_{ij} = \frac{1}{s_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} s_{ki}s_{kj}\right) & j=i+1,\ldots,n\ ; i=2,3,\ldots,n.
\end{cases}
$$

Primele două relaţii se stabilesc cu uşurinţă. Pentru $2 \leq i < j$ au loc egalităţile

$$a_{ij} = \sum_{k=1}^{n} s_{ki}s_{kj} = \sum_{k=1}^{i-1} s_{ki}s_{kj} + s_{ii}s_{ij} + \sum_{k=i+1}^{n} s_{ki}s_{kj}\ .$$

Ultimul termen al sumei este zero deoarece $k > i$ şi prin urmare obţinem ultima dintre relaţiile 3.1(care, pentru $i=j$ dă penultima dintre relaţiile 3.1).

Aceste formule sînt baza pentru <u>metoda rădăcinii pătrate</u> (a lui Banachiewicz) de rezolvare a sistemelor liniare cu matrice simetrică, pozitiv definită (cu minorii diagonali principali pozitivi).

<u>Observaţia 3.2</u>: Ne-am ocupat pînă acum de factorizarea matricelor sub formă de produse de matrice triunghiulare (inferior sau superior) şi matrice diagonale. Desigur această preocupare este legată de găsirea unor procedeie eficiente de rezolvare numerică a sistemelor de ecuaţii liniare. Intr-adevăr, dacă matricea A a sistemului $A\cdot X = b$ se poate descompune sub forma $A = L\cdot U$ cu $L \in I_1^{(n)}$ şi $U \in S^{(n)}$, atunci, rezolvînd sistemul $L\cdot Y = b$ prin substituţie directă şi sistemul $U\cdot X = Y$ prin substituţie inversă, se obţine soluţia sistemului dat.*

Avînd în vedere faptul că sistemele liniare cu matrice ortogonală se puteau rezolva uşor, apare problema factorizării unei matrice pătratice sub forma unui produs dintre o matrice ortogonală şi o matrice triunghiulară (inferior sau superior).

In cele ce urmează ne vom ocupa de această problemă.

Teorema 3.8: Fie $A \in \mathcal{M}_n(R)$ o matrice nesingulară. Atunci:

a) există o matrice $B \in I^{(n)}$ şi o matrice $C \in GO_n(R)$ astfel încît $A = B \cdot C$,

b) există o matrice $B_1 \in S^{(n)}$ şi o matrice $C_1 \in GO_n(R)$ astfel încît $A = C_1 \cdot B_1$. (Aici $GO_n(R)$ notează grupul matricelor ortogonale, de ordinul n, cu elemente din R).

Demonstraţie: a) Presupunem că matricea nesingulară $A \in \mathcal{M}_n(R)$ a fost factorizată sub forma $A = B \cdot C$ cu $B \in I^{(n)}$ şi $C \in GO_n(R)$. Condiţia ${}^tC \cdot C = I_n$ implică

$$^t(B^{-1}A) \cdot (B^{-1}A) = {}^tA \cdot {}^t(B^{-1})B^{-1}A = I_n$$

sau, echivalent

$$^t(B^{-1}) \cdot B^{-1} = {}^t(A^{-1}) \cdot A^{-1} .$$

Rezultă că $A \cdot {}^tA = B \cdot {}^tB$, adică B este transpusa matricei $S \in S^{(n)}$ dintr-o factorizare Cholesky a matricei $A \cdot {}^tA$. Conform teoremei 3.7, existenţa factorizării Cholesky, cu matrice factori reale, este asigurată de condiţia ca minorii diagonali principali ai matricei $A \cdot {}^tA$ să fie pozitivi. Ţinînd cont că submatricea $\begin{pmatrix} 1 & 2 & \dots & k \\ 1 & 2 & \dots & k \end{pmatrix}_{A \cdot {}^tA}$ se obţine ca produs al submatricei A_k formată din primele k linii ale matricei A, cu submatricea ${}^tA_k$ formată din primele k coloane ale matricei ${}^tA$ şi folosind formula Binet-Cauchy, rezultă:

$$\det\begin{pmatrix} 1 & 2 & \dots & k \\ 1 & 2 & \dots & k \end{pmatrix}_{A \cdot {}^tA} = \sum_{1 \le p_1 < \dots < p_k \le n} \det\begin{pmatrix} 1 & 2 & \dots & k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}_{A_k} \cdot \det\begin{pmatrix} p_1 & p_2 & \dots & p_k \\ 1 & 2 & \dots & k \end{pmatrix}_{{}^tA_k} =$$

$$= \sum_{1 \le p_1 < \dots < p_k \le n} \left[\det\begin{pmatrix} 1 & 2 & \dots & k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}_{A_k}\right]^2 > 0$$

(suma este > 0 pentru că, măcar pentru o alegere a indicilor $1 \le p_1 < \dots < p_k \le n$, vom avea $\det\begin{pmatrix} 1 & 2 & \dots & k \\ p_1 & p_2 & \dots & p_k \end{pmatrix}^2_{A_k} \neq 0$).

Minorii diagonali principali ai matricei $A \cdot {}^tA$ fiind > 0, matricea admite o factorizare Cholesky $A \cdot {}^tA = {}^tS \cdot S$ cu $S \in S^{(n)}$, S avînd elemente reale. Punînd atunci $B = {}^tS$, găsim $C = ({}^tS)^{-1} \cdot A$.

Se verifică ușor că ${}^tC\cdot C = I_n$, adică prin acest procedeu am realizat o factorizare de tipul cerut pentru matricea A.

b) Afirmația rezultă din afirmația a) în felul următor. Se știe că matricea nesingulară ${}^tA$ se poate pune sub forma ${}^tA = B\cdot C$, unde $B={}^tS$, $C=B^{-1}\cdot {}^tA$ și ${}^tA\cdot A={}^tS\cdot S$. Rezultă atunci că $A = {}^tC\cdot {}^tB = {}^tC\cdot S$. Luînd $C_1={}^tC \in GO_n(R)$ și $B_1 = S$ obținem o factorizare de tipul indicat.

Ca procedeu practic de lucru indicăm:

- se determină $B_1=S$ din factorizarea Cholesky ${}^tA\cdot A={}^tS\cdot S$,
- se determină $C_1=A\cdot S^{-1}$.

<u>Definiția 3.5</u>: a) <u>Dacă matricea pătratică</u> $A \in \mathcal{M}_n(K)$ <u>a fost pusă sub forma</u> $A=B\cdot C$ <u>cu</u> $B \in I^{(n)}$ <u>și</u> $C \in GO_n(K)$ <u>spunem că am realizat L.Q.-factorizarea (I.O.-factorizarea) matricei</u> A.

b) <u>Dacă matricea pătratică</u> $A \in \mathcal{M}_n(K)$ <u>a fost pusă sub forma</u> $A=C_1\cdot B_1$ <u>cu</u> $C_1 \in GO_n(K)$ <u>și</u> $B_1 \in S^{(n)}$ <u>spunem că am realizat Q.R.-factorizarea (O.S.-factorizarea) matricei</u> A.

Deoarece factorizarea Cholesky nu este unică, nici L.Q.-factorizarea și nici Q.R.-factorizarea nu vor fi unice.

Este de remarcat faptul că în vreme ce L.D.U.-factorizările aveau existența asigurată de satisfacerea condiției K_n, L.Q.-factorizarea și Q.R.-factorizarea se aplică tuturor matricelor nesingulare.

<u>Exemplul 3.6</u>: Considerăm matricea

$$A = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Ne propunem să determinăm L.Q.-factorizarea și Q.R.-factorizarea acestei matrice. Construim

$$A\cdot {}^tA = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}.$$

O factorizare Cholesky a matricei $A\cdot {}^tA$ se realizează cu ajutorul matricei

$$S = \begin{bmatrix} 3 & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & 2\sqrt{\frac{2}{3}} & \sqrt{\frac{2}{3}} \\ 0 & 0 & \sqrt{2} \end{bmatrix}$$

(adică $A \cdot {}^tA = {}^tS \cdot S$).Matricea $B \in I^{(3)}$ este:

$$B = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ -\frac{\sqrt{3}}{3} & 2\sqrt{\frac{2}{3}} & 0 \\ \frac{\sqrt{3}}{3} & \sqrt{\frac{2}{3}} & \sqrt{2} \end{bmatrix}$$

iar matricea $C = B^{-1} \cdot A$ este:

$$C = \frac{1}{4} \begin{bmatrix} \frac{4}{\sqrt{3}} & \frac{4}{\sqrt{3}} & -\frac{4}{\sqrt{3}} \\ 4\frac{\sqrt{6}}{3} & -2\cdot\frac{\sqrt{6}}{3} & 2\frac{\sqrt{6}}{3} \\ 0 & 2\sqrt{2} & 2\sqrt{2} \end{bmatrix} \in GO_3(R) \quad .$$

Matricele B şi C astfel determinate dau o L.Q.-factorizare a matricei A.Pentru determinarea unei Q.R.-factorizări,facem mai întîi factorizarea Cholesky a matricei ${}^tA \cdot A$ .Avem:

$${}^tA \cdot A = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix} \quad .$$

O factorizare Cholesky a matricei ${}^tA \cdot A$ se obţine cu ajutorul matricei:

$$S = \begin{bmatrix} \sqrt{3} & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & 2\frac{\sqrt{6}}{3} & -2\frac{\sqrt{6}}{6} \\ 0 & 0 & \sqrt{2} \end{bmatrix} \quad .$$

Luînd $B_1 = S$ şi $C_1 = A \cdot S^{-1}$, adică

$$C = \begin{bmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{6}}{3} & 0 \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{2} \end{bmatrix} ,$$

se obţine o Q.R.-factorizare a matricei A.*

Aşa cum am observat mai înainte,dacă pentru factorizarea Cholesky lucrăm cu convenţia ca elementele diagonalei principale ale matricei S să fie strict pozitive,atunci factorizarea devine

unică şi tot unice vor fi şi celelalte factorizări legate de factorizarea Cholesky.

Remarcăm faptul că teoremele prezentate anterior rezolvă de obicei existenţa (şi uneori unicitatea) diferitelor tipuri de factorizări. Ar fi de dorit să se prezinte metode efective de realizare a factorizărilor. Menţionăm că procedeie de realizare a L.Q. şi R.Q.-factorizărilor sînt deja cunoscute de la cursul de algebră şi anume: dacă ortonormăm prin procedeul Gram-Schmidt sistemul de vectori coloană ai matricei A realizăm de fapt o L.Q.-factorizare a acestei matrice, iar dacă ortonormăm prin acelaşi procedeu sistemul de vectori linie ai matricei A, realizăm o Q.R.-factorizare a matricei A.

In ultima parte a acestui paragraf ne ocupăm de factorizarea unor matrice dreptunghiulare de rang r. Astfel de factorizări se folosesc pentru construcţia inversei generalizate care, la rîndul ei, se foloseşte la determinarea celei mai bune aproximări (în sensul metodei celor mai mici pătrate) pentru soluţiile unor sisteme de ecuaţii liniare.

Vom arăta mai întîi că o matrice de tip m x n şi rang r se poate factoriza sub forma $A=B\cdot C$ cu $B\in\mathcal{M}_{mxr}(K)$ şi $C\in\mathcal{M}_{rxn}(K)$. Evident, o astfel de factorizare nu este unică, deoarece în cazul în care există o factorizare $A=B_o\cdot C_o$ cu $B_o\in\mathcal{M}_{mxr}(K)$ şi $C_o\in\mathcal{M}_{rxn}(K)$ atunci, pentru orice matrice nesingulară $D\in\mathcal{M}_r(K)$ obţinem factorizările: $A=(B_oD^{-1})(D\cdot C_o) = (B_oD)(D^{-1}C_o)$.

Fie $A\in\mathcal{M}_{mxn}(K)$, cu rang A = r. Presupunem că submatricea $A_{11}=\begin{pmatrix}1 & 2 & \dots & r\\ 1 & 2 & \dots & r\end{pmatrix}_A$ are rangul r. Scriem atunci

$$A = \begin{bmatrix} A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}$$

şi construim matricele

(3.2) $$B = \begin{bmatrix} I_r\\ A_{21}A_{11}^{-1}\end{bmatrix} \qquad C = \begin{bmatrix} A_{11} & A_{12}\end{bmatrix}$$

Vom avea:

$$B\cdot C=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{21}A_{11}^{-1}A_{12}\end{bmatrix} \quad (3.3)$$

Rămîne de dovedit că $A_{22}=A_{21}A_{11}^{-1}A_{12}$. Deoarece rangul matricei A este r, ultimele m-r linii sînt combinaţii liniare ale primelor r linii, adică există o matrice X de tip (m-r)xr astfel încît $X\cdot\begin{bmatrix}A_{11} & A_{12}\end{bmatrix}=\begin{bmatrix}A_{21} & A_{22}\end{bmatrix}$. Rezultă că vor avea loc egalităţile: $X\cdot A_{11}=A_{21}$ şi $X\cdot A_{12}=A_{22}$. Cum $X=A_{21}A_{11}^{-1}$, rezultă $A_{22}=A_{21}A_{11}^{-1}A_{12}$.

Am presupus mai înainte că matricea A_{11} are rangul r. Dacă această condiţie nu este îndeplinită, înmulţind la stînga şi la dreapta cu matrice de tip

$$I_{ij}=\begin{bmatrix}e_1 \dots \underset{i}{e_j} \dots \underset{j}{e_i} \dots e_n\end{bmatrix}$$

obţinute din matricea unitate I_n schimbînd între ele coloanele (liniile) i şi j (aici e_i notează coloana i a matricei unitate I_n), se obţine matricea $A'=PAQ$ pentru care $\begin{pmatrix}1 & 2 & \dots & r\\ 1 & 2 & \dots & r\end{pmatrix}_{A'}$ are determinant nenul. (Reamintim că $I_{ij}A$ este matricea care se obţine din matricea A schimbînd între ele coloanele i şi j, iar AI_{ij} este matricea care se obţine din matricea A schimbînd între ele coloanele i şi j; deci atît P cît şi Q se obţin făcînd produse finite de matrice de tipul I_{ij}, adică P memorează schimbările de linii iar Q - schimbările de coloane necesare obţinerii unui minor de ordin r, nenul, în colţul stînga sus). Pentru A' putem folosi construcţia precedentă şi obţinem factorizarea $A'=B'C'$, cu $B'\in\mathcal{M}_{mxr}(K)$ şi $C'\in\mathcal{M}_{rxn}(K)$. Vom avea atunci $A=PA'Q=(PB')(C'Q)$, adică am obţinut o factorizare de tipul cerut pentru matricea A. (Am folosit faptul că $I_{ij}^{-1}=I_{ij}$ şi prin urmare $P^{-1}=P$ şi $Q^{-1}=Q$.)

Am stabilit următorul rezultat:

<u>Teorema 3.9</u>: <u>Orice matrice</u> $A\in\mathcal{M}_{mxn}(K)$ <u>de rang r poate fi factorizată sub forma</u> $A=B\cdot C$ <u>unde</u> $B\in\mathcal{M}_{mxr}(K)$ <u>şi</u> $C\in\mathcal{M}_{rxn}(K)$.•

<u>Exemplul 3.7</u>: Considerăm matricea

$$A = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 4 & -2 & 0 \\ 2 & 3 & 1 & -1 \end{bmatrix}.$$

Se verifică ușor că rang$A = 2$, dar că $\det\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}_A = \begin{vmatrix} 1 & 2 \\ 2 & 4 \end{vmatrix} = 0$. Considerăm matricea

$$A' = I_{23}A = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 3 & 1 & -1 \\ 2 & 4 & -2 & 0 \end{bmatrix}$$

care are $\det\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}_{A'} = -1$. Impărțim A' în blocuri sub forma $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ unde $A_{11} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$, $A_{12} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$, $A_{21} = \begin{bmatrix} 2 & 4 \end{bmatrix}$ și $A_{22} = \begin{bmatrix} -2 & 0 \end{bmatrix}$8

Rezultă că

$$A_{11}^{-1} = \begin{bmatrix} -3 & 2 \\ 2 & -1 \end{bmatrix} \qquad A_{21}A_{11}^{-1} = \begin{bmatrix} 2 & 0 \end{bmatrix}$$

și că $A' = B' \cdot C'$ cu

$$B' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 0 \end{bmatrix} \qquad C' = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 3 & 1 & -1 \end{bmatrix}.$$

Deci $A = I_{23} \cdot A' = I_{22} \cdot B' \cdot C' = B \cdot C$, unde $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 0 \end{bmatrix}$ și $C = C'$.

<u>Definiția 3.6</u>: <u>Se numește pseudoinversa matricei</u> $A \in \mathcal{M}_{mxn}(K)$ (K este R sau C) <u>matricea notată cu</u> A^I, <u>definită prin relația</u>

(3.4) $$A^I = C^*(C \cdot C^*)^{-1}(B^* \cdot B)^{-1} \cdot B^*$$

<u>unde</u> $A = B \cdot C$ <u>reprezintă o factorizare oarecare a matricei</u> A, <u>cu</u> $B \in \mathcal{M}_{mxr}(K)$ <u>și</u> $C \in \mathcal{M}_{rxn}(K)$, <u>iar</u> B^* (respectiv C^*) <u>este transpusa conjugatei matricei</u> B (respectiv C).

Se știe că factorizarea de tipul indicat în teorema 3.9 nu este unică. Inseamnă că definiția pseudoinversei este corectă numai dacă nu depinde de factorizare. Fie deci $A = B' \cdot C'$, cu $B' \in \mathcal{M}_{mxr}(K)$ și $C' \in \mathcal{M}_{rxn}(K)$. Va exista atunci o matrice $D \in \mathcal{M}_r(K)$ astfel încît $B' = B \cdot D$ și $C' = D^{-1}C$. Au loc egalitățile:

$$C'^*(C'C'^*)^{-1}(B'^*B')^{-1}B'^* = C^*(D^*)^{-1}[D^{-1}(CC^*)D^{*-1}]^{-1}[D^*(B^*B)D]^{-1}B^*B^* =$$
$$= C^*(D^*)^{-1}D^*(CC^*)^{-1}DD^{-1}(B^*B)^{-1}D^{*-1}D^*B^* = C^*(CC^*)^{-1}(B^*B)^{-1}B^*,$$

care arată că pseudoinversa nu depinde de factorizarea pe care o folosim în definiție.

<u>Teorema 3.10</u>: <u>Pseudoinversa</u> $A^I \in \mathcal{M}_{nxm}(K)$ <u>a matricei</u> $A \in \mathcal{M}_{mxn}(K)$ <u>are proprietățile</u>:

$$(3.4)\quad \begin{aligned} &1^\circ.\quad A^I\cdot A\cdot A^I = A^I\\ &2^\circ.\quad A\cdot A^I\cdot A = A\\ &3^\circ.\quad A\cdot A^I = (A\cdot A^I)^*\\ &4^\circ.\quad A^I\cdot A = (A^I\cdot A)^* \end{aligned}$$

Demonstraţia o propunem ca exerciţiu cititorului.

Se poate arăta că proprietăţile 1°– 4° caracterizează pseudoinversa unei matrice. Mai exact, dacă o matrice $B\in\mathcal{M}_{n\times m}(K)$ satisface condiţiile: $B\cdot A\cdot B=B$, $A\cdot B\cdot A = A$, $A\cdot B=(A\cdot B)^*$ şi $B\cdot A=(B\cdot A)^*$ atunci $B=A^I$. Nu justificăm nici această afirmaţie. Acesta este motivul pentru care condiţiile 1°– 4° sînt considerate ca axiomele de definiţie ale pseudoinversei şi sînt cunoscute sub numele de <u>axiomele Moore-Penrose</u>.

Dacă matricea $A\in\mathcal{M}_n(K)$ este nesingulară, atunci $A^I = A^{-1}$.

In teorema 3.8 ne-am ocupat de L.Q.-factorizarea şi Q.R.-factorizarea matricelor pătratice nesingulare. Se poate pune aceeaşi problemă şi pentru matricele dreptunghiulare. In acest caz, putem folosi în locul matricelor ortogonale matrice dreptunghiulare $A\in\mathcal{M}_{m\times n}(K)$ pentru care ${}^tA\cdot A = I_n$ sau $A\cdot {}^tA = I_m$, după cum $m\geqslant n$ sau $m<n$. Prin abuz de limbaj vom numi matricele A cu această proprietate tot matrice ortogonale şi vom vorbi de L.Q.-factorizarea sau Q.R.-factorizarea matricelor dreptunghiulare, după cum $m\leqslant n$ sau $m>n$.

In mod asemănător cu teorema 3.8 se poate demonstra:

<u>Teorema 3.11</u>: a) <u>Pentru orice matrice</u> $A\in\mathcal{M}_{m\times n}(K)$ <u>cu</u> $m\leqslant n$ <u>şi</u> rang $A = m$ <u>există o matrice</u> $B\in I^{(m)}$ <u>şi o matrice</u> $C\in\mathcal{M}_{m\times n}(K)$ <u>cu proprietatea</u> $C\cdot {}^tC = I_m$, <u>astfel încît</u> $A = B\cdot C$.

b) <u>Pentru orice matrice</u> $A\in\mathcal{M}_{m\times n}(K)$ <u>cu</u> $m\geqslant n$ <u>şi</u> rang $A = n$, <u>există o matrice</u> $B_1\in\mathcal{M}_{m\times n}(K)$ <u>astfel încît</u> ${}^tB_1\cdot B_1 = I_n$ <u>şi o matrice</u> $C_1\in S^{(n)}$ <u>astfel încît</u> $A = B_1\cdot C_1$.*

Demonstraţia o propunem ca exerciţiu cititorului. Menţionăm

că B se determină din Factorizarea Cholesky $A\cdot{}^tA = B\cdot{}^tB$,iar C_1 se determină din factorizarea Cholesky ${}^tA\cdot A = {}^tC_1\cdot C_1$ .In plus $C = B^{-1}\cdot A$ şi $B_1 = A\cdot C_1^{-1}$.

<u>Exemlul 3.8</u>: Considerăm matricea

$$A = \begin{bmatrix} 2 & 1 & -1 & 0 \\ 1 & -1 & 0 & 2 \\ 3 & 2 & 1 & -1 \end{bmatrix} .$$

Ne propunem să determinăm L.Q.-factorizarea matricei A.Observăm că rang $A = 3$.Construim $A\cdot{}^tA$

$$A\cdot{}^tA = \begin{bmatrix} 6 & 1 & 7 \\ 1 & 6 & -1 \\ 7 & -1 & 15 \end{bmatrix} .$$

Factorizarea Cholesky a matricei $A\cdot{}^tA$ este $A\cdot{}^tA = B\cdot{}^tB$,unde

$$B = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 1/\sqrt{6} & \sqrt{35/6} & 0 \\ 7/\sqrt{6} & -13/6\cdot\sqrt{6/35} & \sqrt{211/35} \end{bmatrix} .$$

Se verifică uşor că $B\cdot{}^tB = L_3$.In plus

$$C = B^{-1}\cdot A = \begin{bmatrix} \sqrt{2} & 1/\sqrt{6} & -1/\sqrt{6} & 0 \\ 2/3\sqrt{6/35} & -7/6\sqrt{6/35} & 1/6\sqrt{6/35} & 2\sqrt{6/35} \\ -32/35\sqrt{35/211} & 2/5\sqrt{35/211} & 78/35\sqrt{35/211} & -9/35\sqrt{35/211} \end{bmatrix}$$

Teorema 3.11 ne asigură că putem ortonorma un sistem de vectori şi anume sistemul vectorilor linie ai matricei A în cazul $m < n$ şi rang $A = m$ şi respectiv sistemul vectorilor coloană în cazul $m > n$=rang A.Acest procedeu este echivalent cu procedeul Gram-Schmidt.

§4. <u>Procedeul de tridiagonalizare al lui Lanczos</u>

Fie $(E_n; \langle,\rangle)$ un spaţiu euclidean real,n-dimensional şi $T : E_n \longrightarrow E_n$ un operator liniar.Dacă $E_n = E^1 \oplus E^2 \oplus .. \oplus E^s$ este descompunerea în sumă directă de subspaţii invariante (idecompozabile în sumă directă de subspaţii invariante) atunci pentru un vector $w \in E_n$ de forma $w=w_1+...+w_s$,cu $w_i \in E^i$ şi $w_i \neq 0$ pentru orice $i=1,2,...,s$, putem construi familia de vectori $\{w, Tw, T^2w, ..., T^kw, ...\}$ (aici $T^kw = (T\circ T\circ ..\circ T)(w)$).Să presupunem că operatorul liniar T este nesingular.Atunci

$T^k w = T^k w_1 + \dots + T^k w_s$ dă descompunerea lui $T^k w$ în sumă directă și în plus, $T^k w_i \neq 0$ $(i=1,2,\dots,s)$. Vom arăta că familia $\{w, Tw, \dots, T^{n-1}w\}$ este o bază pentru E_n. Intr-adevăr, dacă $T^i w = \sum_{j=0}^{i-1} \alpha_j T^j w$ (cu $T^0 w = w$) atunci subspațiul E', generat de sistemul de vectori $\{w, Tw, \dots, T^{i-1}w\}$ este subspațiu invariant al operatorului T și prin urmare $E' = E^{i_1} \oplus \dots \oplus E^{i_k}$ $(k < s)$. Ar înseamna că $T^i w_j = 0$ pentru $j \notin \{i_1, \dots, i_k\}$, lucru imposibil.

Bazat pe acest rezultat, Lanczos propune următorul procedeu de tridiagonalizare a matricei operatorului T.

Se alege $v_1 \in E_n$ astfel încît proiecția lui pe fiecare subspațiu invariant E^i $(i=1,2,\dots,s)$ să nu fie nulă. Se scrie vectorul Tv_1 sub forma

$$Tv_1 = \langle Tv_1, v_1 \rangle \cdot v_1 + v_2' \quad . \tag{4.1}$$

Se verifică ușor că $v_2' \neq 0$ și că $v_2' \perp v_1$. Relația (4.1) se poate scrie sub forma

$$Tv_1 = \alpha_1 v_1 + \beta_2 v_2 \tag{4.2}$$

unde $\alpha_1 = \langle Tv_1, v_1 \rangle$ iar $\beta_2 = \| Tv_1 - \alpha_1 v_1 \|$; deci v_2 este versorul lui v_2' și prin urmare $\beta_2 = \langle Tv_1, v_2 \rangle$.

Se definește apoi v_{i+1} prin relația de recurență

$$Tv_i = \alpha_i v_i + \gamma_i v_{i-1} + \beta_{i+1} v_{i+1} \tag{4.3}$$

unde $\alpha_i = \langle Tv_i, v_i \rangle$, $\gamma_i = \langle Tv_i, v_{i-1} \rangle$ și $\beta_{i+1} = \| Tv_i - \alpha_i v_i - \gamma_i v_{i-1} \|$ (sau $\beta_{i+1} = \langle Tv_i, v_{i+1} \rangle$). Relația (4.3) ne asigură că $v_{i+1}' = \beta_{i+1} v_{i+1}$ este proiecția ortogonală a vectorului Tv_i pe complementul ortogonal al subspațiului generat de $\{v_{i-1}, v_i\}$.

După efectuarea pasului $n-1$ se obține o bază $\mathcal{B} = \{v_1, v_2, \dots, v_n\}$ cu proprietatea că oricare trei vectori consecutivi $\{v_{i-1}, v_i, v_{i+1}\}$ sînt ortogonali doi cîte doi. De aici nu rezultă neapărat că baza $\mathcal{B}$ este ortonormată. In echimb, matricea operatorului nesingular T, în această bază, este o matrice tridiagonală avînd elementele $a_{ij} \neq 0$ pentru $i-j \in \{0,1\}$ și $a_{ij} = 0$ pen-

tru $|i-j|\geq 2$. Dacă operatorul T este singular, E_n se descompune în suma directă $E_n = E' \oplus E''$, unde $E' = \ker T$ şi E'' este un subspaţiu invariant al lui T, suplementar lui E'. Alegînd o bază ortonormată oarecare în E', matricea operatorului $T_{E'}$ (restricţia lui T la E') este matricea nulă. Pentru restricţia $T_{E''}$ a lui T la E" putem aplica procedeul prezentat mai sus şi ajungem la o bază în care matricea operatorului $T_{E''}$ este tridiagonală.

Dacă în particular, operatorul T este liniar simetric, atunci:

$$\gamma_i = \langle Tv_i, v_{i-1}\rangle = \langle v_i, Tv_{i-1}\rangle = \beta_i \,, \tag{4.4}$$

adică matricea operatorului T în baza $\mathcal{B}$ (construită mai sus) este o matrice tridiagonală simetrică. In acest caz se poate dovedi că baza $\mathcal{B}$ este ortonormată. Intr-adevăr, ştim că primii trei versori din $\mathcal{B}$ sînt ortogonali doi cîte doi. Presupunem că $\{v_1, v_2, \dots, v_i\}$ este un sistem ortonormat de vectori. Pentru orice $1 \leq j \leq i-2$, avem

$$\langle v_j, Tv_i\rangle = \alpha_i \langle v_j, v_i\rangle + \gamma_i \langle v_j, v_{i-1}\rangle + \beta_{i+1}\langle v_j, v_{i+1}\rangle.$$

Conform ipotezei inductive, $\langle v_j, v_i\rangle = \langle v_j, v_{i-1}\rangle = 0$. In plus,

$$\langle v_j, Tv_i\rangle = \langle Tv_j, v_i\rangle = \langle \alpha_j v_j + \gamma_j v_{j-1} + \beta_{j+1} v_{j+1}, v_i\rangle =$$

$$= \alpha_j \langle v_j, v_i\rangle + \gamma_j \langle v_{j-1}, v_i\rangle + \beta_{j+1}\langle v_{j+1}, v_i\rangle = 0,$$

deci $\langle v_j, v_{i+1}\rangle = 0$ pentru $1 \leq j \leq i-2$. Dar prin construcţie v_{i-1} şi v_i sînt ortogonali pe v_{i+1}, deci $\{v_1, v_2, \dots, v_i, v_{i+1}\}$ este un sistem ortonormat de vectori. Prin urmare, în baza $\mathcal{B}$, matricea operatorului liniar simetric T (nesingular) este :

$$[T]_{\mathcal{B}} = \begin{bmatrix} \alpha_1 & \beta_2 & 0 & 0 & \dots & 0 & 0 \\ \beta_2 & \alpha_2 & \beta_3 & 0 & \dots & 0 & 0 \\ 0 & \beta_3 & \alpha_3 & \beta_4 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \alpha_{n-1} & \beta_n \\ 0 & 0 & 0 & 0 & \dots & \beta_n & \alpha_n \end{bmatrix}.$$

Am prezentat mai sus procedeul de tridiagonalizare a matricei unui operator pentru că, în acest caz lucrurile sînt mai

ușor de interpretat. Să ne ocupăm de tridiagonalizarea unei matrice reale simetrice de ordinul n. Vom lucra cu spațiul aritmetic R^n organizat în mod natural ca spațiu euclidean (cu produsul scalar care induce norma euclideană pe R^n). Pentru comoditatea prezentării vom identifica fiecare element al lui R^n cu matricea coloană avînd ca elemente componentele corespunzătoare ale vectorului. Cu această convenție, $\langle v,w\rangle = {}^t v\cdot w$, iar $A\cdot v$ va fi imaginea vectorului v prin operatorul liniar simetric care, în baza canonică din R^n, are matricea A.

Cu aceste precizări, procedura Lanczos de tridiagonalizare a matricei simetrice nesingulare, cu elemente reale, de ordinul n, A, se poate prezenta sub forma:

<u>Pas 1</u> Se alege versorul v_1 generat aleator (pentru a asigura că are proiecții nenule pe fiecare subspațiu invariat al matricei A),

<u>Pas 2</u> Se iau $\beta_0 = 0$, $v_0 = 0$,

<u>Pas 3</u> Pentru i=1,2,...,n se definesc vectorii v_i (vectorii Lanczos) și scalarii α_i și β_{i+1} prin relațiile:

(4.5) $$\beta_{i+1}\, v_{i+1} = A\, v_i - \alpha_i\, v_i - \beta_i\, v_{i-1}$$

(4.6) $$\alpha_i = {}^t v_i\cdot A v_i \quad , \qquad \beta_{i+1} = {}^t v_{i+1}\cdot A v_i .$$

Elementele α_i și β_{i+1} definite prin (4.6) sînt elementele unei matrice tridiagonale $[a_{ij}]$ avînd $a_{ii} = \alpha_i$, $a_{ij} = \beta_{i+1}$ dacă $|i-j| = 1$ și $a_{ij} = 0$ dacă $|i-j| > 1$.

§5. <u>Metode directe de rezolvare a sistemelor de ecuații liniare</u>

<u>Regula lui Cramer</u>

Dacă sistemul

(5.1) $$A\cdot X = b$$

are matricea A pătratică (de ordinul n) și nesingulară, atunci soluția sa are componentele

(5.2) $$x_i = \det A_i \,/\, \det A \qquad i=1,2,\ldots,n$$

unde A_i este matricea obţinută din matricea A prin înlocuirea coloanei i cu coloana termenilor liberi.

Această metodă de rezolvare a sistemelor de ecuaţii liniare nu este de uz practic, ci este de un real folos teoretic. Intr-adevăr, rezolvarea sistemului (5.1) prin regula lui Cramer impune evaluarea a n+1 determinanţi. Evaluarea directă (pe baza definiţiei) a valorii determinantului unei matrice de ordinul n cere efectuarea a $n!(n-1)$ înmulţiri. Deci soluţia sistemului (5.1) se obţine efectuînd $(n+1)!(n-1)$ înmulţiri, a n împărţiri şi a $(n+1)\left[(n+1)!-1\right]$ adunări sau scăderi. Este deci nevoie de un număr foarte mare de operaţii. Pentru a ne da seama de mărimea numărului de operaţii necesare, în condiţiile precizate, pentru rezolvarea sistemului (5.1), prezentăm (după N.Macon) calculul timpului necesar unui calculator care efectuează un milion de operaţii pe secundă, pentru a efectua numai înmulţirile necesare evaluării (pe baza definiţiei) a valorii unui determinant de ordinul 14. Aproximînd $14! \approx 8,72 \cdot 10^{10}$, ţinînd cont că pentru un singur termen avem nevoie de 13 înmulţiri, găsim

$$0,000013 \cdot 8,72 \cdot 10^{10} \text{ sec.} \approx 1.130.000 \text{ sec} \approx 13 \text{ zile}$$

(pentru efectuarea înmulţirilor necesare calculului unui determinant de ordinul 15, acelaşi calculator ar trebui să lucreze circa 212 zile).

Regula lui Cramer se poate folosi cu condiţia să se utilizeze metode mai economice de evaluare a valorii determinanţilor. Un astfel de procedeu i se atribuie lui Chio şi constă în reducerea calculului valorilor determinanţilor unor matrice oarecare la calculul valorilor determinanţilor unor matrice triunghiulare care se obţin din matricele date prin metoda eliminării a lui Gauss. Rezolvarea sistemului (5.1) folosind procedura lui Chio cere circa $\frac{1}{3}n^4$ înmulţiri (un număr total de $\alpha \cdot n^4$ operaţii, cu $\frac{2}{3} < \alpha < 1$).

Metoda lui Gauss

Se aplică de obicei sistemelor Cramer. Fie deci sistemul de ecuaţii liniare

(5.3) $$\sum_{j=1}^{n} a_{ij}\, x_j = b_i \qquad (i=1,2,\ldots,n)$$

cu $\det\left[a_{ij}\right] \neq 0$. Presupunem că $|a_{11}| = \max |a_{i1}|$ (în caz contrar schimbăm ordinea ecuaţiilor în aşa fel încît să avem satisfăcută această condiţie). Se trece apoi la sistemul echivalent:

(5.4) $$\begin{cases} x_1 + a_{12}^{(1)} x_2 + \ldots\ldots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ \qquad a_{22}^{(1)} x_2 + \ldots\ldots + a_{2n}^{(1)} x_n = b_2^{(1)} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad a_{n2}^{(1)} x_2 + \ldots\ldots + a_{nn}^{(1)} x_n = b_n^{(1)} \end{cases}$$

obţinut astfel:

a) se împarte prima ecuaţie la a_{11} şi se obţin coeficienţii $a_{1j}^{(1)} = a_{1j}/a_{11}$ $(1 \leq j \leq n)$,

b) ecuaţia obţinută la a), înmulţită pe rînd cu $-a_{21}, -a_{31}, \ldots, -a_{n1}$ se adaugă respectiv ecuaţiilor de pe locurile $2,3,\ldots,n$ şi se obţin coeficienţii

$$a_{ij}^{(1)} = a_{ij} - a_{i1}\cdot a_{1j}^{(1)}, \qquad b_i^{(1)} = b_i - a_{i1}\cdot b_1^{(1)}.$$

Aplicînd acest procedeu subsistemului format din ecuaţiile (5.4_2) - (5.4_n) se ajunge la un sistem echivalent cu sistemul (5.3), de forma:

(5.5) $$\begin{cases} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots\ldots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ \qquad\qquad x_2 + a_{23}^{(2)} x_3 + \ldots\ldots + a_{2n}^{(2)} x_n = b_2^{(2)} \\ \qquad\qquad\quad a_{33}^{(2)} x_3 + \ldots\ldots + a_{3n}^{(2)} x_n = b_3^{(2)} \\ \qquad\qquad\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad\quad a_{n3}^{(2)} x_3 + \ldots\ldots + a_{nn}^{(2)} x_n = b_n^{(2)} \end{cases}$$

Aplicînd acest procedeu de cel mult încă n-2 ori se ajunge la următorul sistem cu matrice triunghiulară echivalent cu (5.3)

(5.6) $$\begin{cases} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots\ldots + a_{1n}^{(1)} x_n = b_1^{(1)} \\ \qquad\qquad x_2 + a_{23}^{(2)} x_3 + \ldots\ldots + a_{2n}^{(2)} x_n = b_2^{(2)} \\ \qquad\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

$$\left[\begin{array}{l} x_{n-1} + a_{n-1,n}x_n = b_{n-1}^{(n-1)} \\ \qquad\qquad x_n = b_n^{(n)} \end{array}\right. .$$

La etapa k, coeficienţii $a_{ij}^{(k)}$ şi $b_i^{(k)}$ se determină cu ajutorul formulelor:

$$(5.7)\quad a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & i=1,2,..,k-1 \\ a_{kj}^{(k-1)}/\ a_{kk}^{(k-1)} & ,\ j=k,k+1,..,n \\ a_{ij}^{(k-1)} - a_{ik}^{(k-1)}a_{kj}^{(k)} & ,\ i=k+1,...,n;\ j=k,..,n. \end{cases}$$

$$(5.8)\quad b_i^{(k)} = \begin{cases} b_i^{(k-1)} & i=1,2,,,,k-1 \\ b_k^{(k-1)}/\ a_{kk}^{(k-1)} & ,i=k \\ b_i^{(k-1)}-a_{ik}^{(k-1)}b_k^{(k)} & i=k+1,...,n \end{cases} .$$

Sistemul (5.6) se rezolvă apoi prin substituţie inversă (vezi §2,6). Numerele $a_{11}, a_{22}^{(1)},...,a_{ii}^{(i-1)},..,a_{nn}^{(n-1)}$ prin care se împart respectiv ecuaţiile 1,2,..,i,..,n ale sistemelor liniare echivalente care apar (respestiv la etapele 1,2,..,i,..,n) se numesc <u>pivoţi</u>. Pivoţii au fost alegi ca fiind cei mai mari în valoare absolută pe prima coloană a matricei subsistemului de care ne ocupăm la o anumită etapă. Deoarece sistemul (5.3) este sistem Cramer, toţi pivoţii vor fi nenuli. Dacă pivoţii se aleg în acest mod se spune că s-a realizat <u>pivotarea parţială</u>. Se poate alege drept pivot la o anumită etapă elementul cel mai mare în valoare absolută dintre toate elementele matricei subsistemului de care ne ocupăm la acea etapă. In acest caz se spune că se face <u>pivotare totală</u>.

Se poate lucra fără pivotare, cu condiţia să ne asigurăm că, la fiecare etapă k, elementul $a_{kk}^{(k-1)} \neq 0$. Acest lucru este totdeauna posibil de realizat prin schimbarea locului ecuaţiilor. Din păcate, în acest caz, erorile de rotunjire afectează mult rezultatele. In cazul pivotării parţiale sau totale se diminuează influienţa erorilor de rotunjire dar creşte volumul

de lucru.De exemplu,în cazul pivotării totale,pentru a aduce pivotul pe locul (1,1) al matricei subsistemului de interes la o anumită etapă de lucru,trebuiesc efectuate nu numai schimbări de linii,ci şi de coloane; în timp ce schimbările de linii revin la schimbarea ordinei ecuaţiilor sistemului,schimbarea de coloane revine la renumerotarea necunoscutelor sistemului,renumerotare care,în afară de faptul că trebuie realizată, trebuie şi memorată.

Menţionăm că metoda lui Gauss se aplică şi sistemelor de m ecuaţii cu n necunoscute.

Procedeul lui Gauss se poate folosi şi pentru determinarea inversei unei matrice.Se aplică în acest scop procedeul lui Gauss pentru rezolvarea simultană a n sisteme

$$A\cdot X = I_n.$$

De asemeni se poate evalua det A,ca fiind produsul pivoţilor înmulţit cu $(-1)^s$,s fiind numărul total de schimbări de linii şi coloane efectuate în cele n etape de lucru.

Metoda Gauss-Jacobi

Se aplică în aceleaşi condiţii ca şi metoda lui Gauss.

Prima etapă este la fel cu prima etapă a metodei lui Gauss.In a 2-a etapă se trece la sistemul echivalent

$$\begin{cases} x_1 + \qquad + a_{13}^{(2)}x_3 + \dots + a_{1n}^{(2)}x_n = b_1^{(2)} \\ \qquad x_2 + a_{23}^{(2)}x_3 + \dots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \qquad\qquad a_{33}^{(2)}x_3 + \dots + a_{3n}^{(2)}x_n = b_3^{(2)} \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \qquad\qquad a_{n3}^{(2)}x_3 + \dots + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases}$$

folosind transformările:

- se împarte ecuaţia a doua la $a_{22}^{(1)}$
- ecuaţia obţinută se înmulţeşte pe rînd cu $-a_{12}^{(1)},-a_{32}^{(1)},\dots,-a_{n2}^{(1)}$ şi se adaugă respectiv ecuaţiilor 1,3,4,..,n.

Repetînd acest procedeu de n-2 ori se obţine soluţia sistemu-

lui.Evident,metoda se poate folosi pentru evaluarea inversei unei matrice nesingulare şi pentru calculul valorilor determinanţilor.

Metoda rădăcinii pătrate (Cholesky)

Se aplică sistemelor Cramer a căror matrice este simetrică şi pozitiv definită (adică are minorii diagonali principali pozitivi).Folosind formulele (3.1) se obţine factorizarea Cholesky $A={}^tS\cdot S$ cu $S\in S^{(n)}\subset\mathcal{M}_n(R)$.Sistemul $A\cdot X = b$ devine

$$ {}^tS\cdot S\cdot X = b \ . $$

Rezolvînd sistemele

$$ {}^tS\cdot Y = b \qquad\qquad S\cdot X = Y $$

primul prin substituţie directă,iar al doilea prin substituţie inversă,se obţine soluţia sistemului dat.Se apreciază că această metodă este cea mai eficientădintre metodele de rezolvare a sistemelor cu matrice reale simetrice şi pozitiv definite.Anumite probleme le creează cei n-1 radicali care trebuie evaluaţi.

§6.Metode iterative de rezolvare a sistemelor de ecuaţii liniare

O metodă iterativă este o regulă de obţinere a unei soluţii aproximative folosind una sau mai multe soluţii aproximative deja cunoscute.Metodele iterative sînt preferate la rezolvarea sistemelor mari; cu matrice care au multe zerouri (astfel de sisteme apar din aproximaţiile cu diferenţe finite pentru ecuaţii cu derivate parţiale) deoarece ele pot folosi avantajele prezenţei zerourilor matricei.In plus,sînt autocorectoare şi de aceea tind să minimizeze erorile de rotunjire.Astfel de metode sînt bune,în particular,pentru rezolvarea sistemelor cu matrice aproape diagonale sau dominant-diagonale,pentru care convergenţa este rapidă.Ele nu sînt la fel de utile pentru determinarea inversei unei matrice sau pentru rezolvarea simultană a mai multor sisteme de ecuaţii liniare (cu aceeaşi matrice

a coeficienţilor necunoscutelor) deoarece pentru fiecare coloană de termeni liberi are loc separat cîte un proces de convergenţă.De aceea astfel de procedeie se aplică cîte unui singur sistem $A.X = b$.

Uneori convergenţa sistemului este lentă sau este neregulată.Ultima situaţie apare cînd valorile proprii ale matricei A sînt complexe.Convergenţa lentă poate fi remediată folosind procedeie de accelerare.

Dat fiind sistemul Cramer $A.X = b$ (cu $A \in \mathcal{M}_n(R)$ şi $b \in R^n$) definim procedeul iterativ de grad m ca fiind o relaţie de forma

(6.1) $$x^{(k+1)} = G(k,A,b,x^{(k)},x^{(k-1)},\dots,x^{(k-m+1)})$$

unde $x^{(k)}$ notează vectorul cu n componente obţinut la iteraţia numărul k.Se cere ca $x^{(k)} \longrightarrow h = A^{-1}b$ cînd $k \longrightarrow \infty$ (convergenţa folosită fiind de obicei cea corespunzătoare unei norme pe R^n). De obicei m este mic,deoarece în caz contrar ar trebui folosită masiv memoria suplimentară.

Procesu iterativ se numeşte <u>staţionar</u> dacă G este independentă de k,adică $G = G(A,b,x^{(k)},\dots,x^{(k-m+1)})$.Procesul este liniar dacă G este funcţie liniară de $x^{(k)},\dots,x^{(k-m+1)}$.In cazul particular m=1 (aşa-numitele procese Wittmeyer)

(6.2) $$x^{(k+1)} = G(k,A,b,x^{(k)}),$$

orice procedeu iterativ de gradul întîi are una din formele

(6.3) $$x^{(k+1)} = x^{(k)} + D^{(k)}p^{(k)}$$

(6.4) $$x^{(k+1)} = B^{(k)}x^{(k)} + c^{(k)}$$

unde,în general,matricele $D^{(k)}$,$B^{(k)}$ şi vectorii $p^{(k)}$,$c^{(k)}$ sînt funcţii de k,A,b,şi $x^{(k)}$.Se presupune că $D^{(k)}$ şi $B^{(k)}$ sînt nesingulare şi că $D^{(k)} \neq A^{-1}$.

Pentru ca un procedeu iterativ să poată fi utilizat este necesar să lase invariant vectorul soluţie,adică dacă $x^{(k)}=A^{-1}b$ atunci $x^{(k+1)}= A^{-1}b$.Aceasta se întîmplă dacă $D^{(k)}p^{(k)}= 0$ pentru

$x^{(k)} = A^{-1}b$ (adică $P^{(k)} = 0$ pentru $x^{(k)} = A^{-1}b$). şi respectiv $C^{(k)} = (I - B^{(k)})A^{-1}b$. Aici şi în cele ce urmează, vom nota cu I matricea unitate de ordinul n.

Definim <u>k-vectorul rezidual</u> $r^{(k)} = b - Ax^{(k)}$ şi <u>k-vectorul eroare</u> $e^{(k)} = h - x^{(k)}$, unde h este vectorul soluţie. Procedeul iterativ converge dacă şi numai dacă $e^{(k)} \longrightarrow 0$ pentru $k \longrightarrow \infty$. Deoarece h este necunoscut, se folosesc ca teste de convergenţă sau $\|r^{(k)}\|$ sau $\|x^{(k)} - x^{(k-1)}\|$. Ne ocupăm de cazul particular $P^{(k)} = r^{(k)}$. In acest caz $D^{(k)}P^{(k)} = D^{(k)}(b - Ax^{(k)}) = 0$ dacă $x^{(k)} = A^{-1}b$. Relaţia de recurenţă devine

$$x^{(k+1)} = x^{(k)} + D^{(k)}r^{(k)}$$

sau

$$x^{(k+1)} = (I - D^{(k)}A)\, x^{(k)} + D^{(k)}b.$$

Dacă $D^{(k)}$ şi $B^{(k)}$ nu depind de $x^{(k)}$ ci numai de k, A şi b, se obţin <u>iteraţii liniare</u>. Mai exact se obţin relaţiile

(6.5) $$x^{(k+1)} = (I - D^{(k)}A)x^{(k)} + D^{(k)}b$$

şi respectiv

(6.6) $$x^{(k+1)} = B^{(k)}x^{(k)} + (I - B^{(k)})\, A^{-1}b\ ,$$

Scăzînd din (6.5) şi (6.6) egalităţile $A^{-1}b = (I - D^{(k)}A).A^{-1}b + D^{(k)}b$ şi respectiv $A^{-1}b = B^{(k)}x^{(k)} + (I - B^{(k)})A^{-1}b$ se obţin egalităţile

(6.7) $$e^{(k+1)} = (I - D^{(k)}A).e^{(k)}$$

(6.8) $$e^{(k+1)} = B^{(k)}.e^{(k)}$$

Inductiv, din (6.7) şi (6.8) se obţin egalităţile

(6.9) $$e^{(k+1)} = M_{k+1}.e^{(0)}$$

(6.10) $$e^{(k+1)} = M'_{k+1}.e^{(0)}$$

unde

$$M_{k+1} = \prod_{i=0}^{k}(I - D^{(i)}A) \qquad M'_{k+1} = \prod_{i=0}^{k} B^{(i)}$$

sînt matricele de iteraţie (sau de reducere a erorii). Procesul iterativ este convergent dacă şi numai dacă şirurile de matrice $\{M_k\}_{k\in N}$ şi $\{M'_k\}_{k\in N}$ converg la zero cînd $k \longrightarrow \infty$. Dacă toate valorile proprii ale matricelor M_k şi respectiv M'_k sînt mai mici ca 1 în valoare absolută, atunci procesul iterativ conver-

ge. Notăm că, deoarece $r^{(k)} = A\cdot e^{(k)}$ sau $e^{(k)}=A^{-1}r^{(k)}$, au loc egalităţile:

(6.11) $$r^{(k+1)} = (I - AD^{(k)})\cdot r^{(k)}$$

(6.12) $$r^{(k+1)} = (AB^{(k)}A^{-1})\cdot r^{(k)} \quad .$$

Deoarece $AD^{(k)}$ şi $D^{(k)}A$ (respectiv $B^{(k)}$ şi $AB^{(k)}A^{-1}$) au aceleaşi valori proprii, rezultă că criteriile de convergenţă pentru $e^{(k)}\longrightarrow 0$ şi cele pentru $r^{(k)}\longrightarrow 0$, formulate în termenii valorilor proprii ale matricelor de iteraţie, sînt aceleaşi.

Situaţia se simplifică în cazul în care iteraţia este staţionară. In acest caz $D^{(k)}$ şi $B^{(k)}$ nu depind de k şi le vom nota cu D şi respectiv B. Atunci (I - DA) şi B (şi puterile lor) sînt matrice de iteraţie şi, ţinînd cont de (6.9) şi (6.10), găsim:

(6.13) $$e^{(k+1)} = (I - DA)^{k}\cdot e^{(0)}$$

şi respectiv

(6.14) $$e^{(k+1)} = B^{k}\cdot e^{(0)} \quad .$$

Rezultă că, pentru orice normă vectorială $\|\ \|$ şi norma matriceală indusă (notată cu acelaşi simbol $\|\ \|$) vor avea loc inegalităţile:

(6.15) $$\|e^{(k+1)}\| \le \|I-DA\|^{k}\cdot\|e^{(0)}\|$$

şi respectiv

(6.16) $$\|e^{(k+1)}\| \le \|B\|^{k}\cdot\|e^{(0)}\| \quad ,$$

Prin urmare, dacă $\|I-DA\|<1$ (respectiv $\|B\|<1$) atunci procesul iterativ este convergent, indiferent de alegerea lui $x^{(0)}$. Dacă în particular raza spectrală $\rho(I-DA)$ (respectiv $\rho(B)$) este <1, şirul aproximaţiilor succesive este convergent. Din păcate aplicaţia $A\longrightarrow\rho(A)$ ($\rho(A)$ notează raza spectrală a matricei A) nu este o normă matriceală. Totuşi, dacă pentru o matrice A, raza spectrală $\rho(A)<1$, atunci pentru orice $\varepsilon>0$ există o normă matriceală $\|\ \|_{\varepsilon}$ astfel încît

(6.17) $$\|A\|_{\varepsilon}<\rho(A)+\varepsilon \quad .$$

(vezi [10] pag.39). Alegînd ε astfel încît $\rho(A)+\varepsilon<1$, rezultă că în norma $\|\ \|_{\varepsilon}$ şirul aproximaţiilor succesive este convergent.

Cum în R^n toate normele sînt echivalente,rezultă că şirul aproximaţiilor succesive va fi convergent în orice normă (ceea ce va diferi de la o alegere la alta a normei va fi eventual viteza de convergenţă).

Observaţia 6.1 Se apelează la raza spectrală nu numai pentru motivele menţionate mai sus ci şi pentru faptul că există criterii de evaluare rapidă a ordinului de mărime al razei spectrale.(vezi Cap.IV,§2).

Vom prezenta în cele ce urmează un alt criteriu de convergenţă a şirurilor de aproximaţii succesive construite prin procedeie iterative particulare,dar de mare utilitate practică.

Teorema 6.1 Dacă matricea B a procedeului iterativ liniar

(6.18) $$x^{(k+1)} = B\cdot x^{(k)} + C$$

are norma $\|B\| \leq q < 1$, atunci:

1. Sistemul $x=Bx+C$ are soluţia unică ξ,

2. oricare ar fi $x^{(o)} \in R^n$ şirul $\{x^{(k)}\}_{k \in N}$ construit prin formula (6.18) este convergent la ξ,

3. $$\|\xi - x^{(k)}\| \leq \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \leq \frac{q^k}{1-q}\|x^{(1)} - x^{(o)}\|.$$

Demonstraţie Afirmaţiile teoremei rezultă imediat din teorema de punct fix a lui Banach (vezi Cap.I) dacă arătăm că aplicaţia $D : R^n \longrightarrow R^n$,unde $D(x) = Bx + C$ este o contracţie,Intr-adevăr, relaţiile

$$\|D(x)-D(y)\| = \|(Bx+C)-(By+C)\| = \|B(x-y)\| \leq \|B\|\cdot\|x-y\| \leq q\cdot\|x-y\|$$

$(0 \leq q < 1)$ ne asigură că D este o contracţie a spaţiului R^n.Folosind teorema de punct fix a lui Banach,rezultă că au loc afirmaţiile teoremei. □

Corolar 6.2 Dacă matricele I-DA şi B ale proceselor iterative staţionare liniare

$$x^{(k+1)} = (I-DA)x^{(k)} + Db \quad \text{şi} \quad x^{(k+1)} = Bx^{(k)} + (I-B)A^{-1}b$$

au normele $\|I-DA\| \leq q_1 < 1$ şi respectiv $\|B\| \leq q_2 < 1$, atunci pentru sistemele $x=(I-DA)x+Cb$ şi respectiv $x=Bx+(I-B)A^{-1}b$ (echivalente

cu AX = b) au loc afirmaţiile 1°,2° şi 3° ale teoremei 6.1.

Consecinţa 6.3 Dacă matricea B a sistemului liniar $x = Bx + C$ satisface condiţia

(6.19) $$\sum_{j=1}^{n} |b_{ij}| \le \alpha < 1 \qquad (\forall\, i=1,2,\dots,n)$$

atunci, pentru procedeul recursiv

$$x^{(k+1)} = B\,x^{(k)} + C$$

au loc afirmaţiile 1° şi 2° ale teoremei 6.1 şi în plus

(6.20) $$|\xi_i - x_i^{(k)}| \le \frac{\alpha}{1-\alpha} \max_{1\le j\le n} |x_j^{(k)} - x_j^{(k-1)}| \le \frac{\alpha^k}{1-\alpha} \max_{1\le j\le n} |x_j^{(1)} - x_j^{(0)}|,$$

$\forall\, i=1,2,\dots,n$.

Demonstraţie Se observă că $\|B\|_\infty \le \alpha < 1$ şi se aplică teorema 6.1 (şi definiţia normei $\|\cdot\|_\infty$).

Consecinţa 6.4 Dacă matricea B a sistemului liniar $x = Bx + C$ satisface condiţia

(6.21) $$\sum_{i=1}^{n} |b_{ij}| \le \beta < 1 \qquad (\forall\, j=1,2,\dots,n)$$

atunci, pentru procedeul iterativ

$$x^{(k+1)} = Bx^{(k)} + C$$

au loc afirmaţiile 1° şi 2° ale teoremei 6.1 şi în plus

(6.22) $$|\xi_i - x_i^{(k)}| \le \frac{\beta}{1-\beta} \sum_{j=1}^{n} |x_j^{(k)} - x_j^{(k-1)}| \le \frac{\beta^k}{1-\beta} \sum_{j=1}^{n} |x_j^{(1)} - x_j^{(0)}|$$

$\forall\, i=1,2,\dots,n$.

Demonstraţie Se observă că $\|B\|_1 \le \beta < 1$ şi se aplică teorema 6.1 (şi definiţia normei $\|\ \|_1$).

Metode iterative neliniare

Metoda gradientului

Considerăm din nou sistemul

(6.23) $$A\cdot X = b \quad .$$

Construim forma pătratică pozitiv definită asociată unei matrice pătratice simetrice, pozitiv definite $S \in \mathcal{M}_n(R)$

(6.24) $$\phi(x) = {}^t r \cdot S r$$

unde $r = b - Ax$. Se observă că $\phi(x) = 0$ dacă şi numai dacă x este soluţia sistemului (6.23). Deci rezolvarea sistemului (6.23)

este echivalentă cu determinarea vectorului x pentru care $\phi(x) = 0$.

Considerăm procesul iterativ

(6.25) $$x^{(k+1)} = x^{(k)} + \alpha^{(k)} p^{(k)}$$

pentru care $\alpha^{(k)}$ este un scalar, iar $p^{(k)}$ este un vector. Vom determina $\alpha^{(k)}$ şi $p^{(k)}$ pas cu pas în aşa fel încît descreşterea lui $\phi(x)$ să fie cît mai rapidă. Cea mai bună alegere pentru $p^{(k)}$ ar fi desigur direcţia $h-x^{(k)}$ a vectorului eroare dar, din păcate, acest vector nu este cunoscut. Se ştie însă că direcţia în care $\phi(x)$ descreşte cel mai rapid este direcţia gradientului. Deci o alegere potrivită pentru $p^{(k)}$ este $(\mathrm{grad}\,\phi)(x^{(k)})$.

De aceea vom evalua gradientul funcţiei $\phi(x)$. Au loc egalităţile

$$\phi(x) = {}^t r \cdot Sr = {}^t(Sr)\cdot r = \sum_{i=1}^{n}(Sr)_i \cdot r_i = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} S_{ij} r_j\Big)\cdot r_i \quad .$$

Deci

$$\frac{\partial \phi}{\partial x_k} = \sum_{i=1}^{n}\sum_{j=1}^{n}\Big[S_{ij}\Big(r_i\frac{\partial r_j}{\partial x_k} + r_j\frac{\partial r_i}{\partial x_k}\Big)\Big] = 2\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij} r_j \frac{\partial r_i}{\partial x_k}$$

(am folosit din nou faptul că S este simetrică, adică $S_{ij} = S_{ji}$).

Deoarece $r_i = b_i - \sum_{j=1}^{n} a_{ij}x_j$ rezultă că $\partial r_i / \partial x_k = -a_{ik}$, şi prin urmare

$$\frac{\partial \phi}{\partial x_k} = -2\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij} r_j a_{ik} = -2\sum_{i=1}^{n} a_{ik}\Big(\sum_{j=1}^{n} S_{ij} r_j\Big) =$$

$$= -2({}^t A\cdot Sr)_k = -2\big[{}^t(AS)\cdot r\big]_k$$

adică

(6.26) $$\frac{\partial \phi}{\partial x} = -2({}^t A\cdot S)\cdot r \; .$$

Vom lua deci $p^{(k)} = ({}^t A\cdot S)\cdot r^{(k)}$ şi vom determina valoarea lui $\alpha^{(k)}$ din condiţia ca ea să realizeze minimul expresiei $\phi(x+\alpha p)$ pentru x şi p alese arbitrar. Folosind egalităţile

$$\phi(x+\alpha p) = {}^t(b-Ax-\alpha Ap)(Sb-SAx-\alpha SAp) = {}^t(r-\alpha Ap)(Sr-\alpha SAp) =$$
$$= {}^t r\cdot Sr - \alpha\,{}^t r\cdot SAp - \alpha\,{}^t(Ap)\cdot Sr + \alpha^2\,{}^t(Ap)S(Ap) = \phi(x) - 2\alpha\,{}^t r\cdot SAp +$$
$$+ \alpha^2\,{}^t(Ap)\cdot S\cdot(Ap),$$

obţinem

(6.27) $$\phi(x+\alpha p) = \phi(x) + \Big(\alpha - \frac{{}^t r\cdot SAp}{{}^t(Ap)SAp}\Big)^2\big[{}^t(Ap)SAp\big] - \Big(\frac{{}^t r\cdot SAp}{{}^t(Ap)\cdot SAp}\Big)^2 .$$

Vom obţine valoare minimă, indiferent de p şi x, atunci cînd

$$(6.28) \qquad \alpha = \frac{{}^t r \cdot SAp}{{}^t(Ap) \cdot SAp} \, .$$

Vom lua deci

$$(6.29) \qquad p^{(k)} = {}^t(SA) \cdot r^{(k)}$$

$$(6.30) \qquad \alpha^{(k)} = \frac{{}^t r^{(k)} \cdot SAp^{(k)}}{{}^t(Ap^{(k)}) \cdot SAp^{(k)}} \, .$$

Am obţinut deci procesul iterativ (6.25) cu $\alpha^{(k)}$ şi $p^{(k)}$ definiţi prin (6.29) şi respectiv (6.30), pentru care

$$\phi(x^{(k+1)}) < \phi(x^{(k)}).$$

Egalitate se obţine numai dacă ${}^t r^{(k)} \cdot SAp^{(k)} = 0$, adică ${}^t[{}^t(SA) r^{(k)}] \cdot [{}^t(SA)\, r^{(k)}] = 0$. Ultima relaţie are loc numai cînd $r^{(k)} = 0$, adică $x^{(k)}$ este soluţia sistemului (6.23). Rămîne să dovedim că $\phi(x^{(k)}) \longrightarrow \hat{0}$ cînd $k \longrightarrow \infty$. Vom dovedi acest rezultat în cazul particular $S = A^{-1}$ (deci A este matrice simetrică şi pozitiv definită).

In acest caz

$$(6.31) \qquad p^{(k)} = r^{(k)} \qquad \alpha^{(k)} = \frac{{}^t r^{(k)} \cdot r^{(k)}}{{}^t r^{(k)} \cdot Ar^{(k)}} \, .$$

Facem raportul

$$\frac{\phi(x^{(k+1)})}{\phi(x^{(k)})} = 1 - \frac{{}^t r^{(k)} \cdot r^{(k)}}{({}^t r^{(k)} \cdot Ar^{(k)})({}^t r^{(k)} \cdot A^{-1} r^{(k)})} \, .$$

Putem folosi acum inegalitatea lui Kantorovici

$$1 \leq \frac{({}^t x \cdot Ax)({}^t x \cdot A^{-1} x)}{{}^t x \cdot x} \leq \left(\frac{\kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}}}{2}\right)^2 \qquad \forall\, x \neq 0,$$

unde $\kappa = \kappa_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2$ este numărul de condiţionare al matricei A (în cazul normei euclidiene). Rezultă că:

$$\frac{\phi(x^{(k+1)})}{\phi(x^{(k)})} \leq 1 - \frac{4}{(\kappa^{\frac{1}{2}} + \kappa^{-\frac{1}{2}})^2} = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 ;$$

deci şirul $\{\phi(x^{(k)})\}_{k \in N}$ converge la zero cel puţin la fel de rapid ca progresia geometrică de raţie $\left(\frac{\kappa - 1}{\kappa + 1}\right)^2$. Prin urmare, κ mare înseamnă convergenţă lentă (şi totodată rea condiţionare a matricei A ; vezi §8).

Observaţia 6.2 1. Drept S se obişnuieşte să se ia

a) $S = (A \cdot {}^t A)^{-1}$ b) $S = I$.

Corespunzător se obţine:

a) $\operatorname{grad}\phi(x) = -2A^{-1}r$, $p = A^{-1}r$, $\alpha = \frac{{}^t(A^{-1}r)\cdot p}{{}^tp\cdot p}$

b) $\operatorname{grad}\phi(x) = -2\,{}^tAr$, $p = {}^tAr$, $\alpha = \frac{{}^tr\cdot p}{{}^tp\cdot Ap}$.

2. Funcţia $\phi(x)$ este cea care dă o măsură globală a erorii pentru diversele alegeri ale lui S.

Metoda gradienţilor conjugaţi

Metodele de direcţii conjugate se bazează pe generarea unei mulţimi A-ortogonale de vectori şi apoi minimizarea succesivă în direcţia fiecăruia dintre ei.

Reamintim că mulţimea $\{p^{(k)}\}$, K=0,1,2,..,n-1 se numeşte A-ortogonală sau A-conjugată dacă ${}^t(Ap^{(i)})\cdot p^{(j)} = 0$ pentru $i\neq j$. Dacă A este pozitiv definită atunci ${}^t(Ap^{(i)})\cdot p^{(i)} > 0$. In acest caz vectorii $p^{(k)}$ sînt liniar independenţi şi, prin urmare, orice soluţie h a sistemului Ax=b poate fi scrisă sub forma

$$h = \sum_{i=0}^{n-1}\alpha_i\cdot p^{(i)} . \tag{6.32}$$

Deoarece $p^{(k)}$ este A-ortogonală, rezultă că

$${}^t(Ah)\cdot p^{(k)} = {}^tb\cdot p^{(k)} = \sum_{i=0}^{n-1}\alpha_i\,{}^t(Ap^{(i)})\cdot p^{(k)} = \alpha_k\,{}^t(Ap^{(k)})\cdot p^{(k)},$$

adică

$$\alpha_k = \frac{{}^t(Ah)\cdot p^{(k)}}{{}^t(Ap^{(k)})\cdot p^{(k)}} = \frac{{}^tb\cdot p^{(k)}}{{}^t(Ap^{(k)})\cdot p^{(k)}} . \tag{6.33}$$

Deci, în momentul în care s-a determinat familia $\{p^{(k)}\}$ de vectori A-ortogonali, se poate obţine soluţia h în n pagi, printr-o iteraţie de forma

$$\begin{aligned} x^{(i+1)} &= x^{(i)} + \alpha_i p^{(i)} \qquad (i=0,1,\dots,n-1)\\ \alpha_i &= \frac{{}^t(Ah)\cdot p^{(i)}}{{}^t(Ap^{(i)})\cdot p^{(i)}} = \frac{{}^tb\cdot p^{(i)}}{{}^t(Ap^{(i)})\cdot p^{(i)}} \end{aligned} \tag{6.34}$$

(i=0,1,..,n-1) începînd cu $x^{(0)} = 0$.

Pentru a determina o familie de vectori A-ortogonali $\{p^{(k)}\}$ (k=0,1,..,n-1), procedăm inductiv în modul următor. Considerăm o bază oarecare $\{v^{(k)}\}_{k=0}^{n}$ şi folosim relaţia de recurenţă

$$(6.35)\qquad \begin{cases} p^{(0)} = v^{(0)} \\ p^{(k+1)} = v^{(k+1)} + \sum_{i=0}^{k} c_i^{(k+1)} p^{(i)} \qquad (k=0,1,..,n-2). \end{cases}$$

Presupunem că $\{p^{(0)},\ldots,p^{(k)}\}$ sînt A-ortogonali. Impunem condiția ${}^t(Ap^{(k+1)}).p^{(r)} = 0$ $(r=0,1,\ldots,k)$ și găsim

$$0 = {}^t(Ap^{(k+1)}).p^{(r)} = {}^t(Av^{(k+1)}).p^{(r)} + \sum_{i=0}^{k} c_i^{(k+1)}\,{}^t(Ap^{(i)}).p^{(r)}$$

pentru $r=0,1,\ldots,k$, adică

$$(6.36)\qquad c_r^{(k+1)} = -\frac{{}^t(Av^{(k+1)}).p^{(r)}}{{}^t(Ap^{(i)}).p^{(r)}} .$$

Se obține astfel o metodă iterativă numită metoda <u>direcțiilor conjugate</u>.

Să trecem acum să prezentăm metoda <u>gradienților conjugați</u>. Această metodă combină metoda direcțiilor conjugate cu metoda gradientului. Este de fapt un caz special al metodei direcțiilor conjugate și anume cazul în care $p^{(k)}$ se obțin prin A-ortogonalizarea vectorilor $r^{(k)}$. O metodă de direcții conjugate în care vectorii reziduali sînt ortogonali între ei este o metodă de gradienți conjugați.

Presupunem că sistemul $Ax = b$ are matricea A simetrică și pozitiv definită. Alegem un vector inițial $x^{(0)}$ și punem

$$p^{(0)} = r^{(0)} = b - Ax^{(0)} .$$

Definim

$$(6.37)\qquad \begin{cases} x^{(1)} = x^{(0)} + \alpha_0 p^{(0)} \\ r^{(1)} = b - Ax^{(1)} \\ p^{(1)} = r^{(1)} + \beta_0 p^{(0)} . \end{cases}$$

Pe α_0 îl determinăm ca soluție a problemei : să se determine valoarea parametrului α pentru care $\phi(x+\alpha p)$ este minimă oricare ar fi x și p (aici $\phi(x) = {}^tr.A^{-1}r$, cu $r=b-Ax$). Ținînd cont de rezultatele din prima parte, găsim

$$(6.38)\qquad \alpha_0 = {}^tp^{(0)}.r^{(0)}/[{}^tp^{(0)}.Ap^{(0)}] .$$

Pe β_0 îl determinăm din relația

$${}^t(Ap^{(1)}).p^{(0)} = 0.$$

Se obține:

(6.39) $$\beta_0 = -[{}^t r^{(1)} \cdot Ap^{(0)}]/[{}^t p^{(0)} \cdot Ap^{(0)}] .$$

Presupunem că am determinat $x^{(k)}, r^{(k)}, p^{(k)}$ astfel încît $\{p^{(0)},\dots, p^{(k)}\}$ să fie A-ortogonali. Din condiția de minimizare pentru $\phi(x)$ găsim

(6.40) $$\alpha_k = [{}^t r^{(k)} \cdot p^{(k)}]/[{}^t p^{(k)} \cdot Ap^{(k)}] .$$

Cu ajutorul lui α_k determinăm $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ și $r^{(k+1)} = b - Ax^{(k+1)}$. Determinăm apoi $p^{(k+1)} = r^{(k+1)} + \beta_k p^{(k)}$, unde β_k este ales astfel încît ${}^t(Ap^{(k)}) \cdot p^{(k+1)} = 0$. Se găsește

(6.41) $$\beta_k = -[{}^t r^{(k+1)} \cdot Ap^{(k)}]/[{}^t p^{(k)} \cdot Ap^{(k)}] .$$

Soluția sistemului va fi dată de $h = \sum_{i=0}^{n-1} \gamma_i p^{(i)}$, unde

$$\gamma_i = \frac{{}^t b \cdot p^{(k)}}{{}^t(Ap^{(i)}) \cdot p^{(k)}} \qquad (i=0,1,\dots,n-1).$$

(vezi formulele (6.34)).

Metode iterative liniare

Vom prezenta numai două metode iterative liniare : metoda lui Jacobi și metoda Gauss-Seidel.

Metoda lui Jacobi

Dat sistemul Cramer de n ecuații cu n necunoscute $Ax = b$, putem pune totdeauna matricea sa A sub forma

$$A = D - J - S$$

unde $D = \mathrm{diag}\,(a_{11}, a_{22}, \dots, a_{nn})$ este matrice nesingulară și

$$-J = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ a_{21} & 0 & 0 & \dots & 0 & 0 \\ a_{31} & a_{32} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n-1,1} & a_{n-1,2} & a_{n-1,3} & & 0 & 0 \\ a_{n,1} & a_{n,2} & a_{n,3} & & a_{n,n-1} & 0 \end{bmatrix}, \quad -S = \begin{bmatrix} 0 & a_{12} & a_{13} & \dots & a_{1,n-1} & a_{1,n} \\ 0 & 0 & a_{23} & \dots & a_{2,n-1} & a_{2,n} \\ 0 & 0 & 0 & \dots & a_{3,n-1} & a_{3,n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & a_{n-1,n} \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Cu aceste notații sistemul se poate scrie sub forma $x = Bx + C$, unde $B = D^{-1}(J+S)$ și $C = D^{-1}b$. Se definește atunci relația de recurență :

(6.42) $$x^{(k+1)} = B\,x^{(k)} + C .$$

Are loc:

<u>Teorema 6.5</u> <u>Dacă matricea A a sistemului Ax=b satisface condiția</u>

(6.43)
$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \qquad (i=1,2,\dots,n)$$

atunci:

1. sistemul $Ax=b$ are soluţie unică (pe care o notăm cu ξ),

2. oricare ar fi $x^{(0)} \in R^n$, şirul $\{x^{(k)}\}$ definit prin relaţia (6.42) este convergent la ξ,

3. $$|\xi_i - x_i^{(k)}| \leq \frac{\alpha}{1-\alpha} \max_{1\leq j\leq n} |x_j^{(k)} - x_j^{(k-1)}| \leq \frac{\alpha^k}{1-\alpha} \max |x_j^{(1)} - x_j^{(0)}|$$

unde α este un număr real care îndeplineşte condiţiile
$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \leq \alpha < 1 \qquad (i=1,2,\dots,n).$$

Demonstraţie Se aplică Consecinţa 6.3. □

Procedeul lui Jacobi poate fi descris în felul următor: Presupunînd cunoscut $x^{(k)} = {}^t[x_1^{(k)},\dots,x_n^{(k)}]$, definim $x^{(k+1)}$ prin

1. înlocuim $x_2, x_3, \dots, x_n$ respectiv cu $x_2^{(k)}, x_3^{(k)}, \dots, x_n^{(k)}$ în prima ecuaţie şi, rezolvînd-o în raport cu x_1, găsim $x_1^{(k+1)}$,

2. înlocuim $x_1, x_3, \dots, x_n$ respectiv cu $x_1^{(k)}, x_3^{(k)}, \dots, x_n^{(k)}$ în a doua ecuaţie şi, rezolvînd-o apoi în raport cu x_2, găsim $x_2^{(k+1)}$,

....

n. înlocuim $x_1, x_2, \dots, x_{n-1}$ respectiv cu $x_1^{(k)}, \dots, x_{n-1}^{(k)}$ în a n-a ecuaţie şi, rezolvînd-o apoi în raport cu x_n, găsim $x_n^{(n+1)}$.

Evident, acest proces se poate realiza numai dacă toate elementele diagonalei principale sînt nenule. Această condiţie este satisfăcută dacă are loc (6.43).

Metoda Gauss-Seidel

Să considerăm din nou sistemul Cramer de n ecuaţii cu n necunoscute $Ax = b$. Folosind notaţiile introduse în secţiunea precedentă, îl punem sub forma echivalentă
$$(D - J)x = Sx + b.$$
Dacă D este nesingulară, atunci D-J este tot nesingulară şi obţinem sistemul echivalent
$$x = (D - J)^{-1} \cdot Sx + (D - J)^{-1} b \quad .$$
Acest sistem este de forma $x=Bx+C$, unde

(6.44) $$B = (D - J)^{-1}S, \qquad C = (D - J)^{-1}b .$$

Are loc

Teorema 6.6 Dacă matricea A a sistemului $Ax = b$ satisface condiția

(6.45) $$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \qquad (i=1,2,..,n)$$

atunci

1. sistemul $Ax=b$ are soluție unică (pe care o notăm cu ξ);
2. oricare ar fi $x^{(o)} \in R^n$, șirul $\{x^{(k)}\}_{k\in N}$ definit prin relația de recurență $x^{(k+1)} = Bx^{(k)} + C$ (cu B și C definiți prin relațiile (6.44)) este convergent la ξ;
3. au loc inegalitățile

$$|\xi_i - x_i^{(k)}| \leq \frac{\alpha}{1-\alpha} \max |x_j^{(k)} - x_j^{(k-1)}| \leq \frac{\alpha^k}{1-\alpha} \max |x_j^{(1)} - x_j^{(o)}| ,$$

$(i=1,2,..,n)$, unde α este un număr real care îndeplinește condițiile

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| \leq \alpha < 1 \qquad (i=1,2,..,n).$$

Demonstrație Se observă că este suficient să se arate că $\|B\|_\infty \leq \alpha < 1$. Pentru aceasta este suficient să arătăm că $\|Bx\|_\infty \leq \alpha \|x\|_\infty$, $\forall x \in R^n$. Facem notația $Bx=Y$, adică $(D-J)Y=Sx$ sau, echivalent, $DY=JY+Sx$. Scriind ultima relație cu ajutorul elementelor matricei A obținem:

$$a_{ii}y_i = -\sum_{j=1}^{i-1} a_{ij}y_j - \sum_{j=i+1}^{n} a_{ij}x_j \qquad (i=1,2,..,n).$$

Pentru $i=1$ se obține $a_{11}y_1 = -\sum_{j=2}^{n} a_{1j}x_j$, sau, echivalent

$$y_1 = \sum_{j=2}^{n} (-a_{1j}/a_{11})x_j .$$

Deci
$$|y_1| \leq \sum_{j=2}^{n} \left|\frac{a_{1j}}{a_{11}}\right| |x_j| \leq \sum_{j=2}^{n} \left|\frac{a_{1j}}{a_{11}}\right| \max_{1\leq j\leq n} |x_j| \leq \alpha \|x\|_\infty .$$

Presupunem că $|y_j| \leq \alpha \|x\|_\infty$ pentru $1 \leq j \leq i-1$. Atunci

$$a_{ii}y_i = -\sum_{j=1}^{i-1} a_{ij}y_j - \sum_{j=i+1}^{n} a_{ij}x_j$$

sau
$$|y_i| \leq \sum_{j=1}^{n-1} \left|\frac{a_{ij}}{a_{ii}}\right| |y_j| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| |x_j| \leq \left(\alpha \sum_{j=1}^{i-1} \left|\frac{a_{ij}}{a_{ii}}\right| + \sum_{j=i+1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|\right) \cdot \|x\|_\infty \leq$$
$$\leq \left(\sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|\right) \|x\|_\infty .$$

Deci, procedînd inductiv, rezultă că $|y_i| \leq \|x\|_\infty$ $(i=1,2,..,n)$, adică

$$\|Bx\|_{\infty} = \|y\|_{\infty} \leq \alpha \|x\|_{\infty}$$

şi prin urmare, $\|B\|_{\infty} \leq \alpha < 1$. Aplicînd Consecinţa 6.3, va rezulta că afirmaţiile teoremei sînt adevărate. □

Şi procedeul Gauss-Seidel poate fi descris în mod asemănător cu procedeul lui Jacobi şi anume:

Dacă $x^{(k)}$ a fost evaluat, determinăm $x^{(k+1)}$ în felul următor:

1. înlocuim $x_2, x_3, \ldots, x_n$ respectiv cu $x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}$ în prima ecuaţie şi, rezolvînd-o apoi în raport cu x_1, găsim $x_1^{(k+1)}$;

2. înlocuim în a doua ecuaţie x_1 cu $x_1^{(k+1)}$ şi $x_3, \ldots, x_n$ respectiv cu $x_3^{(k)}, \ldots, x_n^{(k)}$ şi, rezolvînd-o apoi în raport cu x_2, găsim $x_2^{(k+1)}$;

....

i. înlocuim în ecuaţia i pe $x_1, \ldots, x_{i-1}$ respectiv cu $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ şi $x_{i+1}, \ldots, x_n$ respectiv cu $x_{i+1}^{(k)}, \ldots, x_n^{(k)}$ şi, rezolvînd-o apoi în raport cu x_i, găsim $x_i^{(k+1)}$;

....

n. înlocuim în a n-a ecuaţie $x_1, \ldots, x_{n-1}$ respectiv cu $x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)}$ şi, rezolvînd-o apoi în raport cu x_n, găsim $x_n^{(k+1)}$.

<u>Observaţia 6.3</u> Este plauzibil să credem că metoda Gauss-Seidel este mai rapidă decît metoda lui Jacobi, deoarece pentru calculul componentei $x_i^{(k+1)}$ foloseşte deja componentele calculate $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$. Există totuşi exemple care arată că aceste metode nu sînt comparabile. Mai exact, pentru anumite matrice A raza spectrală a matricei $D^{-1}(J+S) = I - D^{-1}A$ este mai mică decît cea a matricei $(D-J)^{-1}S$, iar pentru altă matrice are loc situaţia opusă. In sprijinul acestei afirmaţii prezentăm un exemplu datorat lui L. Collatz.

<u>Exemplul 6.1</u> Considerăm sistemele $Ax=b$ corespunzătoare matricelor

a) $A = \begin{bmatrix} 1 & -2 & 2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{bmatrix}$ b) $A = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 1 & -1 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$.

Atunci

a) $D^{-1}(J+S)=\begin{bmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{bmatrix}$, şi $(D-J)^{-1}S = \begin{bmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{bmatrix}$.

Prima matrice are polinomul caracteristic $-\lambda^3$, iar a doua $-\lambda(\lambda^2+4\lambda-4)$. Deci razele spectrale ale acestor două matrice sînt 0 şi respectiv $2(1+\sqrt{2})$. Deci, în vreme ce procesul Jacobi converge, procesul Gauss-Seidel nu converge.

b) $D^{-1}(J+S)=\begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 1 \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$ şi $(D-J)^{-1}S = \begin{bmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & -\frac{1}{2} & 3/2 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix}$.

Polinoamele caracteristice ale acestor matrice sînt $-\lambda(\lambda^2+\frac{5}{4})$ şi repectiv $-\lambda(\lambda+\frac{1}{2})^2$. Deci prima matrice are raza spectrală $\frac{\sqrt{5}}{2}$, iar cea de a doua $\frac{1}{2}$. Inseamnă că în acest caz procesul Jacobi nu converge în vreme ce procesul Gauss-Seidel este convergent.

<u>Observaţia 6.4</u> In general, procesul iterativ Jacobi este lent convergent. De aceea se pune problema modificării acestui proces în aşa fel încît să se obţină un proces iterativ mai rapid convergent. Un astfel de procedeu este cunoscut sub denumirea de <u>procedeul Jacobi extrapolat</u> şi se obţine în modul următor:

cunoscînd $x^{(k)}$, se determină mai întîi $\hat{x}^{(k+1)}$ prin

$$\hat{x}^{(k+1)} = D^{-1}(J+S)\, x^{(k)} + D^{-1}b \tag{6.46}$$

şi apoi se determină $x^{(k+1)}$ prin

$$x^{(k+1)} = x^{(k)} + \omega(\hat{x}^{(k+1)} - x^{(k)}), \tag{6.47}$$

ω fiind un parametru ce urmează a fi determinat.

Punînd (6.46) sub forma

$$\hat{x}^{(k+1)} = (I - D^{-1}A)x^{(k)} + D^{-1}b \tag{6.48}$$

şi eliminînd $\hat{x}^{(k+1)}$ din relaţiile (6.47) şi (6.48) se obţine:

$$x^{(k+1)} = (I - \omega D^{-1}A)\, x^{(k)} + \omega D^{-1}b. \tag{6.49}$$

Se constată imediat că procesul iterativ (6.49) lasă invariantă soluţia sistemului $Ax=b$.

Vom face cîteva observaţii cu privire la alegerea lui ω. Parametrul ω se va alege în aşa fel încît raza spectrală

$\rho(I-\omega D^{-1}A)$ să fie mai mică decît $\rho(I-D^{-1}A)$.Fie μ o valoare proprie a matricei $I-\omega D^{-1}A$.Atunci

$$0 = \det[I-\omega D^{-1}A-\mu I] = \det\left[(1-\mu)I-\omega D^{-1}A\right] = $$
$$=\omega^n \det\left[(I-D^{-1}A) - \frac{\mu+\omega-1}{\omega} I\right]$$

şi, prin urmare, $\tilde{\mu} = \frac{\mu+\omega-1}{\omega}$ va fi valoare proprie pentru $I-D^{-1}A$.In mod asemănător se arată că dacă λ este valoare proprie pentru $D^{-1}A$, atunci $\tilde{\mu} = 1-\lambda$ şi $\mu = 1-\omega\lambda$ sînt valori proprii pentru $I-D^{-1}A$ şi respectiv pentru $I-\omega D^{-1}A$.Dacă în plus matricea A este reală,simetrică şi pozitiv definită,tot aşa va fi şi matricea $D^{-1}AD$ şi,prin urmare,valorile proprii ale matricei $D^{-1}A$ vor fi reale şi pozitive.Procesul Jacobi asociat matricei A va converge dacă $|1-\lambda|<1$,adică $\lambda \in (0,2)$,iar procesul (6.49) va converge dacă $\omega\lambda \in (0,2)$.Se apreciază că cea mai bună alegere a lui ω este aceea pentru care $1-\lambda_{min}\omega = -(1-\lambda_{max}\cdot\omega)$,adică

(6.50) $$\omega = \frac{2}{\lambda_{min}+\lambda_{max}} .$$

In acest caz

$$\mu = |1-\lambda\omega| = \left|1 - \frac{2\lambda}{\lambda_{min}+\lambda_{max}}\right| \leq \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}} < 1,$$

deci,cu cît λ_{min} şi λ_{max} sînt mai apropiate, μ este mai mic şi procesul iterativ (6.49) este mai rapid convergent.Formula (6.50) nu este de prea mare utilitate practică,deoarece determinarea lui λ_{max} şi λ_{min} este o problemă greu de rezolvat în multe cazuri.

Să ne ocupăm de problema similară relativă la procesul Gauss-Seidel.Se apreciază că în majoritatea cazurilor practice raza spectrală a matricei de iteraţie a procesului iterativ Gauss-Seidel este de ordinul $1-\mathcal{O}(n^{-2})$,adică şi în acest caz convergenţa este lentă.Definim un nou proces iterativ,legat de procesul Gauss-Seidel,în felul următor:

cunoscînd $x^{(k)}$,se determină mai întîi $\tilde{x}^{(k+1)}$ prin

(6.51) $$D\tilde{x}^{(k+1)} = b + Jx^{(k+1)} + Sx^{(k)}$$

şi apoi se determină $x^{(k+1)}$ prin

(6.52) $$x^{(k+1)} = x^{(k)} + \omega(\hat{x}^{(k+1)} - x^{(k)}),$$

ω fiind un parametru care urmează să fie determinat. Eliminînd $\hat{x}^{(k+1)}$ din relațiile (6.51) și (6.52) se obține

(6.53) $$x^{(k+1)} = (D - \omega J)^{-1}\left[(1-\omega)D + \omega S\right]x^{(k)} + \omega(D - J)^{-1}b.$$

Procesul iterativ (6.53) este cunoscut sub numele de <u>metoda de relaxare</u>. O condiție necesară pentru convergența metodei de relaxare este $0 < \omega < 2$ (nu justificăm această afirmație). Dacă $\omega \in (0,1)$ metoda corespunzătoare este metodă de <u>subrelaxare</u>, iar dacă $\omega \in (1,2)$ este <u>metodă de suprarelaxare</u>. Metodele de suprarelaxare au fost analizate de Young și Frankel. S-a stabilit că în o mare parte de cazuri practice, raza spectrală a matricei de iterație este de ordinul lui $1-\mathcal{O}(n^{-1})$. J. Todd apreciază că metoda lui Young (de suprarelaxare) este de aproape patru ori mai rapidă decît metoda Gauss-Seidel.

Incheiem cu observația că metoda de obținere a proceselor (6.49) și (6.53) se încadrează în următoarea schemă de obținere a unui proces iterativ pornind de la un proces iterativ cunoscut. Fie procesul iterativ $x^{(k+1)} = G(x^{(k)})$; construim procesul iterativ :

$$\hat{x}^{(k+1)} = G(x^{(k)})$$
$$x^{(k+1)} = x^{(k)} + \omega(\hat{x}^{(k+1)} - x^{(k)}),$$

unde ω este un parametru convenabil ales. Există și așa-numitele metodele de accelerare liniară Cebîșev, construite prin:

$$\hat{x}^{(k+1)} = G(x^{(k)})$$
$$x^{(k+1)} = \hat{x}^{(k+1)} + \omega(\hat{x}^{(k+1)} - x^{(k)}).$$

§7. <u>Soluțiile generalizate ale unui sistem de ecuații liniare</u>

Considerăm sistemul $Ax = b$, cu $A \in \mathcal{M}_{mxn}(\mathbb{R})$. Pentru fiecare $x \in \mathbb{R}^m$ construim vectorul reziduu $r(x) = b - Ax$. Dacă x_n este soluție a sistemului, atunci $r(x_n) = 0$; dacă sistemul nu admite soluție se pune problema să determinăm elementul (sau elementele) $x \in \mathbb{R}^n$ pentru care $\|r(x)\|_2$ (norma euclideană) este minimă.

Soluţia acestei probleme se va numi soluţie generalizată sau soluţie în sensul metodei celor mai mici pătrate.

In cele ce urmează vom prezenta un procedeu de determinare a soluţiei generalizate de normă minimă.

Se ştie că soluţia sistemului $Ax = b$, în cazul în care A este matrice pătratică nesingulară, este $x = A^{-1}b$. Apare problema: ce interpretare are $x_n = A^Ib$ (pentru definiţia matricei A^I vezi §2) ? Dovedim mai întîi următoarea lemă:

<u>Lema 7.1</u> <u>Pentru orice</u> $x \in R^n$, <u>au loc egalităţile</u>

(7.1) $${}^t(Ax - AA^Ib)(AA^Ib - b) = 0,$$

(7.2) $${}^t(A^Ib)(x - A^IAx) = 0.$$

<u>Demonstraţie</u> 1) Folosind formulele (3.4) şi proprietăţile operaţiilor cu matrice, se obţin egalităţile:

$${}^t(Ax-AA^Ib)(AA^Ib-b)={}^t(Ax)(AA^Ib)-{}^t(Ax)b-{}^t(AA^Ib)(AA^Ib)+$$
$$+{}^t(AA^Ib)b={}^tx\,{}^tA\,{}^t(AA^I)b-{}^t(Ax)b-{}^tb\,{}^t(AA^I)(AA^I)b+{}^t(AA^Ib)b={}^tx\,{}^t(AA^IA)b-$$
$$-{}^t(Ax)b-{}^tbA(A^IAA^I)b+{}^t(AA^Ib)b={}^tx\,{}^tAb-{}^t(Ax)b-{}^tb(AA^Ib)+{}^t(AA^Ib)b=$$
$$={}^t(Ax)b-{}^t(Ax)b-{}^t(AA^Ib)b+{}^t(AA^Ib)b = 0.$$

2) In acelaşi mod se obţin egalităţile:

$${}^t(A^Ib)(x-A^IAx)={}^t(A^Ib)x-{}^t(A^Ib)(A^IAx)={}^t(A^Ib)x-{}^tb\,{}^t(A^I)\,{}^t(A^IA)x=$$
$$={}^t(A^Ib)x - {}^t(A^Ib)x = 0.$$ □

Folosind rezultatele acestei leme şi norma euclideană rezultă egalităţile

$$\|Ax-b\|^2= \|A(x-A^Ib) + (I-AA^I)(-b)\|^2=\|Ax-AA^Ib\|^2+\|AA^Ib-b\|^2$$

($\forall x \in \mathbb{R}^n$). De aici rezultă

(7.3) $$\|Ax - b\|>\|AA^Ib - b\| = \|Ax_n - b\|$$

cu excepţia acelor x pentru care $Ax - AA^Ib = 0$, adică a acelor x pentru care $Ax = Ax_n$ (deci x este soluţie generalizată). Din această inegalitate rezultă că dacă x_1 este soluţie exactă a sistemului, atunci şi x_0 va fi soluţie exactă. Folosind din nou lema 7.1 se obţine egalitatea:

(7.4) $$\|A^I b + (I - A^I A)x\|^2 = \|A^I b\|^2 + \|(I - A^I A)x\|^2 \ , \forall x \in R^n.$$

Dacă $Ax = Ax_0$ (adică x este soluţie generalizată) atunci

$$A^I Ax = A^I Ax_0 = A^I A A^I b = A^I b \ .$$

Egalitatea (7.4) devine

$$\|x\|^2 = \|x_0\|^2 + \|(I - A^I A)x\|^2 \ .$$

Am stabilit deci următorul rezultat:

Teorema 7.1 Dat sistemul $Ax = b$ cu $A \in \mathcal{M}_{mxn}(R)$ şi A^I inversa generalizată a matricei A, în general $x_0 = A^I b$ este soluţia unică determinată de metoda celor mai mici pătrate. Dacă există x astfel încît $Ax = Ax_0$ atunci x este soluţie generalizată. In acest caz x_0 este caracterizată ca fiind soluţia generalizată cu norma cea mai mică (x_0 este unică cu această proprietate).

§8. Sisteme bine condiţionate şi sisteme rău condiţionate Măsuri ale condiţionării

Cînd erori "mici" în coeficienţi sau în termenii liberi sau în procesul de rezolvare (de exemplu rotunjirile) a unui sistem de ecuaţii liniare au efect "mic" asupra soluţiei, sistemul se numeşte "bine condiţionat"; dacă efectul asupra soluţiei este mare, adică soluţia este foarte sensibilă la variaţii ale valorilor coeficienţilor, se spune că sistemul este "rău condiţionat". După cum sistemul este bine condiţionat sau rău condiţionat, se spune că matricea sistemului este bine condiţionată, respectiv rău condiţionată.

Practica arată că buna sau reaua condiţionare a unei matrice depinde de problema în care apare matricea, adică o matrice poate fi bine condiţionată în raport cu problema rezolvării sistemelor liniare de ecuaţii, dar rău condiţionată în raport cu problema determinării valorilor proprii. Tot practica a arătat că toate metodele de rezolvare a sistemelor rău condiţionate sînt în general rele : metodele directe produc erori de rotunjire semnificative care provoacă dificultăţi în alegerea pivoţilor,

iar metodele iterative converg lent. Aceasta este motivul pentru care rezolvarea problemei condiţionării are o importanţă deosebită. Din păcate, chiar definiţia bunei (respectiv relei) condiţionări este vagă; de fapt nu avem o definiţie (în sensul propriu) a bunei condiţionări, ci numai o descriere a sa.

Se pot face următoarele aprecieri.

Reaua condiţionare în raport cu rezolvarea sistemelor liniare apare atunci cînd liniile ei sînt "aproape" liniar dependente (sau, în limbaj geometric, cînd unele dintre hiperplanele definite de ecuaţiile sistemului fac între ele unghiuri mici adică sînt "aproape" paralele). Totuşi folosirea lui $\det A$ ca măsură a condiţionării nu este indicată, deoarece ecuaţiile pot fi amplificate prin constante care permit obţinerea oricărei valori pentru determinantul matricei sistemului. Cînd ecuaţiile sînt normalizate, adică liniile i $(1 \leq i \leq n)$ sînt împărţite prin $(\sum_{j=1}^{n} a_{ij}^{2})^{2}$ atunci determinantul matricei A poate fi folosit ca test pentru reaua condiţionare. Mai exact, dacă valoarea absolută a determinantului matricei normalizate A este mică în comparaţie cu 1, sistemul este rău condiţionat. Ca exemplu să luăm sistemul

$$\begin{cases} x_1 + 10\,x_2 = 11 \\ 10\,x_1 + 101\,x_2 = 111, \end{cases}$$

a cărui soluţie este $x_1=1$, $x_2=1$. Determinantul matricei A a coeficienţilor necunoscutelor este $\det A = 1$. Valorile proprii ale lui A sînt $\lambda_1 = 101{,}9902$ şi $\lambda_2 = 0{,}0098$. Pentru a normaliza sistemul, împărţim prima ecuaţie la $\sqrt{101}$ şi a 2-a la $\sqrt{10301}$. Atunci determinantul matricei normalizate este $\frac{1}{\sqrt{101}} \cdot \frac{1}{\sqrt{10301}} \simeq 0{,}001$, adică sistemul este rău condiţionat.

Pentru un sistem rău condiţionat, o mulţime de valori date necunoscutelor, care diferă considerabil de soluţie, poate da reziduuri mici. Intr-adevăr, deoarece vectorul eroare e este legat de vectorul rezidual r prin relaţia $e = A^{-1}r$, rezultă

că r mic nu implică în mod necesar că e este mic. Pentru sistemul de mai sus avem: ${}^t[x_1 \ \ x_2] = {}^t[0,95 \ \ 1,05]$ dă ${}^tr = {}^t[0,45 \ \ 4,55]$ adică $\|r\|^2 = 20,90$, în timp ce ${}^t[x_1 \ \ x_2] = [11,1 \ \ 0]$ dă ${}^tr = -[0,1 \ \ 0]$ şi $\|r\|^2 = 0,01$. Se vede că pentru al doilea set de valori ale necunoscutelor, restul este mai mic decît pentru primul set, dar primul set este mult mai aproape de soluţia sistemului. Deci un vector reziduu egal cu zero implică implică faptul că sistemul de valori dat necunoscutelor este soluţie, dar un vector rezidual mic nu înseamnă că sistemul de valori ales este aproape de soluţie. Un rezidual mic înseamnă că nu există erori substanţiale în direcţia proprie corespunzătoare valorii proprii mai mari (în valoare absolută) dar ele pot fi substanţiale în direcţia proprie corespunzătoare valorilor proprii mici. Cînd o matrice are toate valorile proprii apropiate una de alta, adică $|\lambda_{max}|/|\lambda_{min}| \simeq 1$, matricea este bine condiţionată pentru rezolvarea sistemelor liniare (dar este rău condiţionată pentru rezolvarea problemei valorilor proprii). Pentru exemplul de mai sus $|\lambda_{max}|/|\lambda_{min}| = \frac{101,9902}{0,0098} = 10407$.

De obicei se folosesc următoarele măsuri ale condiţionării unei matrice A de ordinul n :

1. Numărul de condiţionare $M(A)$ (al lui Turing) :

$$M(A) = n \cdot \max_{i,j}|a_{ij}| \cdot \max_{i,j}|\alpha_{ij}| \tag{8.1}$$

unde $A = [a_{ij}]$ şi $A^{-1} = [\alpha_{ij}]$; o valoare medie pentru $M(A)$ este $\sqrt{n} \cdot \lg n$.

2. Numărul de condiţionare $N(A)$ (al lui Turing)

$$N(A) = n^{-1}\|A\| \cdot \|A^{-1}\| \tag{8.2}$$

unde $\|A\| = (\sum_{i,j} a_{ij}^2)^{\frac{1}{2}}$; o valoare medie pentru $N(A)$ este $\sqrt{n}$.

3. Numărul de condiţionare $P(A)$ (al lui von Neumann) şi Goldstine):

$$P(A) = \left|\frac{\lambda_{max}}{\lambda_{min}}\right| \tag{8.3}$$

unde λ_{max} şi λ_{min} sînt cea mai mare şi respectiv cea mai mică valoare proprie (în valoare absolută); o valoare medie pentru $P(A)$ este n.

4. Determinantul matricei normalizate. Aşa cum am menţionat mai sus, normalizarea matricei A se obţine împărţind fiecare linie i $(1 \leq i \leq n)$ a matricei $A = [a_{ij}]$ prin $(\sum_{j=1}^{n} a_{ij}^2)^{\frac{1}{2}}$. Dacă determinantul matricei normalizate este mic în comparaţie cu 1, sistemul este rău condiţionat.

Au loc următoarele relaţii

$$\frac{M(A)}{n^2} \leq N(A) \leq M(A) , \tag{8.4}$$

$$\frac{M(A)}{n} \leq P(A) \leq n \cdot M(A). \tag{8.5}$$

Cel mai bine sînt condiţionate matricele ortogonale, care au $N(A) = 1$ şi $M(A)$ în jurul lui logn.

Observaţia 8.1 O. Taussky a dovedit că $P(A \cdot {}^tA) \geq P(A)$ şi $N(A \cdot {}^tA) \geq N(A)$, deci simetrizarea unei matrice rău condiţionate conduce la matrice mairău condiţionate. In concluzie, nu este avantajos să se înlocuiască sistemul $Ax = b$ cu sistemul echivalent ${}^tA \cdot Ax = {}^tA \cdot b$.

Observaţia 8.2 Numerele de condiţionare nu sînt de uz practic, deoarece cer cunoaşterea inversei sau a valorilor proprii, sau determinantul matricei, cantităţi pe care în general dorim să le determinăm. Importanţa cel puţin teoretică a utilizării numerelor de condiţionare $M(A)$ şi $N(A)$ rezultă şi din următoarele două motive:

a) Presupunem că o schimbare Δb a menmbrului drept al sistemului $Ax = b$ cauzează o modificare Δx a soluţiei. Atunci

$$Ax = b \qquad\qquad A(x + \Delta x) = b + \Delta b$$

implică $A \cdot \Delta x = \Delta b$, adică $\Delta x = A^{-1} \cdot \Delta b$. Deci

$$\|\Delta x\| = \|A^{-1} \cdot \Delta b\| \leq \|A^{-1}\| \cdot \|\Delta b\| = \|A^{-1}\| \frac{\|\Delta b\|}{\|b\|} \cdot \|Ax\| \leq \|A\| \|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|b\|} \cdot \|x\| ,$$

adică

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|b\|} .$$

b) Presupunem că modificarea ΔA a matricei A a sistemului $Ax = b$ determină o modificare Δx a soluţiei, adică

$$(A + \Delta A)(x + \Delta x) = b.$$

Rezultă că $A \cdot \Delta x + \Delta A \cdot x + \Delta A \cdot \Delta x = 0$, adică

$$\Delta x = -(A+\Delta A)^{-1} \cdot \Delta A \cdot x = -[A(I + A^{-1} \cdot \Delta A)]^{-1} \cdot \Delta A \cdot x .$$

Se obţine atunci egalitatea

$$\Delta x = -(I + A^{-1} \cdot \Delta A)^{-1} \cdot A^{-1} \cdot \Delta A \cdot x$$

care implică

$$\|\Delta x\| \leq (1 + \|A^{-1}\| \cdot \|\Delta A\| + \|A^{-1}\|^2 \cdot \|\Delta A\|^2 + \ldots) \cdot \|A^{-1}\| \cdot \|\Delta A\| \cdot \|x\|$$

sau

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta A\|}{\|A\|}}{1 - \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta A\|}{\|A\|}}$$

dacă $\|A^{-1}\| \cdot \|\Delta A\| < 1$.

§9. Exerciţii

1. (Kahan) Rezolvaţi sistemul $Ax = b$

a) exact

b) folosind calculatorul

unde

$$A = \begin{bmatrix} 0,2161 & 0,1441 \\ 1,2969 & 0,8648 \end{bmatrix} \qquad b = \begin{bmatrix} 0,1440 \\ 0,8642 \end{bmatrix}$$

c) Calculaţi reziduul $r = b - Az$, unde $z = {}^t[0,9911 \quad -0,4870]$

d) determinaţi A^{-1}

e) calculaţi $\|A\|_1, \|A^{-1}\|_1, \|A\|_\infty$ şi $\|A^{-1}\|_\infty$.

2. (Wilson) a) Rezolvaţi sistemul $Wx = b$ cînd

$$W = \begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix}$$

şi ${}^tb = [32 \quad 23 \quad 33 \quad 31]$ şi ${}^tb = [32,01 \quad 22,99 \quad 33,01 \quad 30,99]$.

b) Folosind factorizarea Cholesky a matricei W, arătaţi că

$$W^{-1} = \begin{bmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{bmatrix} .$$

c) Găsiţi $\|W\|_1, \|W\|_\infty, \|W^{-1}\|_1$ şi $\|W^{-1}\|_\infty$.

3. Este rău condiționat sistemul

$$\begin{cases} 6x + 13y - 17z = 1 \\ 13x + 29y - 38z = 2 \\ -17x - 38y + 50z = -3 \end{cases} \quad ?$$

4. Calculați $\varkappa_{\infty}(A)$ și $\varkappa_{\infty}(DA)$ cînd

$$A = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 10 & 10^2 \\ 1 & 10^2 & 10^4 \end{bmatrix} \qquad D = \operatorname{diag}\left(\frac{1}{3}, \frac{1}{111}, \frac{1}{10101}\right).$$

Ce se poate spune, utilizînd cele două numere de condiționare, despre sistemele $AX = b$ și $DAX = Db$?

5. Găsiți vectorii reziduali $r^{(k)}$ (explicit) în cazul rezolvării sistemului

$$\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} X = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

prin metoda lui Jacobi și prin metoda lui Gauss-Seidel.

6. Rezolvați sistemul

$$\begin{cases} 0,34622x+0,35381y+0,36518z=0,24561 \\ 0,89318x+0,90274y+0,91143z=0,62433 \\ 0,22431x+0,23642y+0,24375z=0,17145 \end{cases}$$

a) folosind metoda lui Gauss cu pivotare parțială

b) folosind metoda lui Gauss cu pivotare totală

c) evaluați rezidualii corespunzători aproximațiilor

$$\begin{bmatrix} -1,027066 \\ 2,091962 \\ -0,380515 \end{bmatrix} \qquad \begin{bmatrix} -1,022773 \\ 2,084125 \\ -0,376941 \end{bmatrix} \qquad \begin{bmatrix} -1,031229 \\ 2,099457 \\ -0,383879 \end{bmatrix}.$$

7. Determinați pseudoinversa matricei

$$A = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}.$$

8. Care este pseudoinversa unei matrice coloană?

9. Transformați sistemul $AX = b$, unde

$$A = \begin{bmatrix} 3 & 10 & -1 & 2 \\ 20 & 4 & 5 & 3 \\ 19 & 5 & -5 & -1 \\ -1 & 1 & -2 & 5 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 15 \\ 21 \\ 17 \end{bmatrix}$$

într-un sistem echivalent, de forma $X = BX + c$ astfel încît $\|B\|_m <$ < 1. Folosind metoda aproximațiilor succesive, pornind de la $x^{(0)} = {}^t[0 \quad 0 \quad 0 \quad 0]$, determinați soluția sitemului cu o eroare $< 10^{-2}$.

10. Folosind metoda lui Jacobi, să se aproximeze soluția sistemului $AX = b$ cu o eroare $\leq 10^{-4}$ cînd

a) $$A = \begin{bmatrix} 10 & 2 & 2 & 1 \\ 2 & 10 & -1 & 1 \\ 2 & 1 & 10 & 1 \\ 1 & 1 & 1 & 10 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

b) $$A = \begin{bmatrix} -3 & 7 & -1 & 2 \\ 14 & 4 & 5 & 3 \\ 11 & 5 & -4 & -1 \\ -1 & 1 & -2 & 7 \end{bmatrix} \qquad b = \begin{bmatrix} 20 \\ 36 \\ 28 \\ 22 \end{bmatrix}.$$

11. Rezolvați sistemele de la exercițiul 10, cu ajutorul metodei Gauss-Seidel (tot cu eroarea $< 10^{-4}$) și comparați atît rezultatele cît și volumul de lucru necesar.

B. REZOLVAREA NUMERICA A ECUATIEI $f(x) = 0$

§1. Generalităţi

Considerăm funcţia $f:D\subset R \longrightarrow R$. Prin soluţie a ecuaţiei

$$f(x) = 0 \qquad (1.1)$$

înţelegem un element $x_0 \in D$ cu proprietatea $f(x_0) = 0$. Obţinerea unei soluţii x_0 a ecuaţiei (1.1) nu este în general posibilă. De aceea, una dintre problemele de bază ale calculului numeric este obţinerea unor aproximaţii acceptabile ale soluţiilor (reale) ale ecuaţiei (1.1).

Dacă $f(x)$ este un polinom, ecuaţia (1.1) se va numi <u>ecuaţie algebrică</u>; în caz contrar ecuaţia (1.1) se va numi <u>ecuaţie transcendentă</u>.

Problema propusă mai sus se referă la determinarea soluţiilor reale ale ecuaţiei (1.1). Pentru ecuaţiile algebrice este important să determinăm şi rădăcinile complexe, adică rădăcinile ecuaţiei $f(z) = 0$ $(z \in C)$. In cele ce urmează ne ocupăm de determinarea / aproximarea rădăcinilor reale ale ecuaţiei (1.1), dar vom face şi scurte aprecieri legate de rădăcinile complexe ale ecuaţiilor algebrice.

Prima problemă pe care ne-o punem în cazul ecuaţiei (1.1) este determinarea numărului de soluţii reale ale ei. Această problemă este intim legată de problema separării soluţiilor.

A <u>separa</u> (<u>a localiza</u>) rădăcinile reale ale ecuaţiei (1.1) înseamnă a determina o familie de intervale $\{I_k\}_{k\in K\subset N}$ disjuncte două cîte două şi cu proprietatea că fiecare interval I_k $(k\in K)$ conţine o soluţie şi numai una singură a ecuaţiei (1.1).

In rezolvarea numerică a ecuaţiei (1.1) distingem două etape importante:

1) separarea rădăcinilor

2) determinarea aproximativă a rădăcinii şi evaluarea erorii care s-a comis.

In cele ce urmează ne ocupăm pe scurt de fiecare dintre aceste probleme.

§2. Localizarea rădăcinilor

Vom prezenta pe scurt cîteva procedeie de separare a rădăcinilor unei ecuaţii (1.1). Menţionăm că metodele obişnuite de separare a rădăcinilor se aplică funcţiilor care au proprietatea lui Darboux.

Fie $I \subset R$ - un interval şi $f: I \longrightarrow R$ o funcţie cu proprietatea lui Darboux. Dacă pentru $a, b \in I$ (cu $a < b$) $f(a) \cdot f(b) < 0$, atunci $f(a)$ şi $f(b)$ au semne contrare; dacă $f(a) < 0$ şi $f(b) > 0$, atunci, cum $f([a,b])$ este un interval, rezultă că $0 \in f([a,b])$, adică există cel puţin un element $x_0 \in (a,b)$ astfel încît $f(x_0) = 0$. La fel se raţionează şi dacă $f(a) > 0$ şi $f(b) < 0$. Dacă în plus f este strict monotonă pe (a,b), ecuaţia va avea în intervalul (a,b) o soluţie unică. Bazaţi pe aceste consideraţii putem dovedi:

<u>Teorema 2.1</u> <u>Fie</u> $I \subset R$ <u>un interval şi</u> $f: I \longrightarrow R$ <u>o funcţie derivabilă. Atunci:</u>

a) <u>Dacă</u> $a, b \in I$, $a < b$, $f(a) \cdot f(b) < 0$ <u>şi</u> $f'(x) \neq 0$ $\forall x \in (a,b)$, <u>atunci ecuaţia</u> $f(x) = 0$ <u>are în intervalul</u> (a,b) <u>o soluţie unică.</u>

b) <u>Dacă</u> f <u>este continuă pe</u> I, <u>derivabilă pe</u> $\mathring{I} = \mathrm{Int}\, I$, <u>cu proprietatea că există</u> $m_1, M_1 > 0$ <u>astfel încît</u> $m_1 \leq |f'(x)| \leq M_1$, $\forall x \in \mathring{I}$, <u>atunci</u>

$$\frac{1}{M_1}|f(x_1) - f(x_2)| \leq |x_1 - x_2| \leq \frac{1}{m_1}|f(x_1) - f(x_2)|, \forall x_1, x_2 \in I. \tag{2.1}$$

<u>Demonstraţie</u> a) f fiind derivabilă pe I, este continuă, deci are proprietatea lui Darboux. Condiţia $f'(x) \neq 0$, $\forall x \in (a,b)$ asigură că pe acest interval funcţia este strict monotonă. Rezultă deci existenţa unei soluţii unice $x_0 \in (a,b)$ pentru ecuaţia $f(x) = 0$.

b) Folosind teorema creşterilor finite pe orice subinterval

$[x_1,x_2]\subset I$, rezultă că există $\xi\in(x_1,x_2)$ astfel încît $f(x_1)-f(x_2)=$ $=f'(\xi)\cdot(x_1-x_2)$. Cum $m_1\le|f'(\xi)|\le M_1$, rezultă că

$$m_1 \le \frac{|f(x_1)-f(x_2)|}{|x_1-x_2|} \le M_1$$

adică are loc relaţia (2.1).

In cazul particular cînd x_0 este soluţie a ecuaţiei (1.1), în condiţiile Teoremei 2.1 b) va rezulta că

(2.2) $$|x-x_0|\le\frac{1}{m_1}|f(x)| \qquad \forall x\in I.$$

Un procedeu practic des întrebuinţat pentru separarea rădăcinilor se bazează pe şirul lui Rolle.

<u>Definiţia 2.1</u> <u>Fie</u> $I\subset R$ <u>un interval de capete</u> $a,b\in\bar{R}$ ($\bar{R}$ <u>este dreapta încheiată</u>) şi $f:I\longrightarrow R$ <u>o funcţie derivabilă astfel încît</u> f' <u>se anulează într-un număr finit de puncte</u> $x_1<x_2<\ldots<x_{n-1}$. <u>Atunci şirul</u> $f(a+),f(x_1),\ldots,f(x_{n-1}),f(b-)$ <u>se numeşte şirul lui Rolle asociat funcţiei</u> f.

Folosind teorema 2.1 se poate dovedi

<u>Teorema 2.2</u> <u>Fie</u> $I\subset R$ <u>un interval de capete</u> $a,b\in R$ <u>şi</u> $f:I\longrightarrow R$ <u>o funcţie derivabilă astfel încît</u> f' <u>se anulează într-un număr finit de puncte</u> $x_1<x_2<\ldots<x_{n-1}$. <u>Dacă există</u> $i\in$ $\in\{1,2,\ldots,n-1\}$ <u>astfel încît</u> $f(x_{i-1})\cdot f(x_i)<0$ (<u>respectiv</u> >0) <u>atunci între</u> x_{i-1} <u>şi</u> x_i <u>se află o singură soluţie (resp. nu se află nici-o soluţie) a ecuaţiei</u> $f(x)=0$ (<u>aici am notat</u> $f(x_0)=f(a+)$ <u>şi</u> $f(x_n)=f(b-)$).

In aceleaşi condiţii ca în Teorema 2.2, numărul de variaţii de semn din şirul lui Rolle din care am suprimat termenii nuli, este egal cu numărul soluţiilor reale ale ecuaţiei $f(x)=0$ situate în mulţimea $(a,b)\setminus\{x_1,x_2,\ldots,x_{n-1}\}$.

Folosirea şirului lui Rolle este condiţionată de posibilitatea determinării rădăcinilor derivatei $f'(x)$ pe intervalul de interes. Cum rezolvarea ecuaţiei $f'(x)=0$ nu este totdeauna posibilă, rezultă că nu putem apela totdeauna la acest procedeu de

separare a rădăcinilor.

Un alt procedeu de separare a rădăcinilor se bazează pe şirul lui Sturm.

Definiţia 2.2 Fie $I \subset R$ un interval şi $f:I \longrightarrow R$ o funcţie derivabilă. Se numeşte şir Sturm asociat lui f un şir finit de de funcţii continue $f_0, f_1, .., f_s : I \longrightarrow R$ cu proprietăţile:

1. $f_0 \equiv f$,
2. $f_s(x) \neq 0$, $\forall x \in I$,
3. dacă există $c \in I$ şi $i \in \{1,2,..,s-1\}$ astfel încît $f_i(c) = 0$, rezultă că $f_{i-1}(c) \cdot f_{i+1}(c) < 0$,
4. dacă există $c \in I$ cu $f(c) = 0$, rezultă $f_0'(c) \cdot f_1(c) > 0$.

Observaţia 2.1 Dacă $f_0, f_1, .., f_s$ este un şir Sturm asociat lui f, atunci $f_0, \lambda_1 f_1, .., \lambda_s f_s$ este tot şir Sturm asociat lui f dacă $\lambda_1, .., \lambda_s$ sînt numere reale strict pozitive.

Fie $f:I \longrightarrow R$ o funcţie derivabilă şi $f_0, f_1, .., f_s$ un şir Sturm asociat funcţiei f. Atunci, pentru un $x \in I$ oarecare, vom nota cu $V(x)$ numărul de variaţii de semn din şirul $f_0(x), f_1(x), ..., f_s(x)$ din care am suprimat elementele nule. De asemenea, pentru $[a,b] \subset I$ vom nota cu $N(a,b)$ numărul rădăcinilor reale din intervalul (a,b) ale ecuaţiei $f(x) = 0$.

Are loc

Teorema 2.3 Fie $I \subset R$ un interval şi $f:I \longrightarrow R$ o funcţie de clasă C^1 care admite un şir Sturm $f_0, f_1, .., f_s$. Atunci oricare ar fi a, b cu $a < b$ şi $f(a) \cdot f(b) \neq 0$ rezultă că numărul rădăcinilor din intervalul (a,b) ale ecuaţiei $f(x) = 0$ este $V(a) - V(b)$ (adică $N(a,b) = V(a) - V(b)$).

Nu prezentăm demonstraţia acestei teoreme. Ne vom ocupa de cazul în care f(x) este un polinom.

Teorema 2.4 (Sturm) Fie $P \in R[X]$ - un polinom cu coeficienţi reali, care nu posedă rădăcini multiple şi fie $P_0, P_1, .., P_s$ un şir finit de polinoame definit astfel : $P_0 = P$, $P_1 = P'$, P_i este

restul cu semn schimbat al împărţirii lui P_{i-2} la P_{i-1} pentru orice $i \in \{2,3,..,s\}$;ultimul polinom P_s fiind o constantă nenulă.Atunci $P_o,P_1,..,P_s$ este un şir Sturm asociat lui P,deci pentru orice $a,b \in R, a<b$,cu $P(a) \neq 0$ şi $P(b) \neq 0$,numărul rădăcinilor reale din intervalul (a,b) ale ecuaţiei $P(x)=0$ este egal cu $V(a) - V(b)$.

Demonstraţie Prin definiţie $P_o = P$ şi $P_1 = P'$.Pentru $i \geq 2$ avem

$$P = P' \cdot q_1 - P_2$$

(2.3) $$P' = P_2 \cdot q_2 - P_3$$

$$..............$$

$$P_{i-2} = P_{i-1} \cdot q_{i-1} - P_i$$

$$..................$$

Cum gradele polinoamelor P_i $(i=0,1,2,..)$ descresc cel puţin cu cîte o unitate,rezultă că există numai un număr finit de polinoame P_i construit în acest mod.Fie P_s ultimul polinom obţinut în acest fel.Am obţinut şirul $\{P_o,P_1,..,P_s\}$ .Cum P nu are rădăcini multiple,rezultă că P_s este o constantă nenulă.Vom dovedi că şirul astfel obţinut este un şir Sturm.Primele două condiţii din definiţia şirului Sturm sînt evident satisfăcute.Dacă $P_i(c) = 0$ atunci relaţia $P_{i-1} = P_i \cdot q_i - P_{i+1}$ conduce la egalitatea $P_{i-1}(c) = -P_{i+1}(c)$ adică $P_{i-1}(c) \cdot P_{i+1}(c) < 0$.In sfîrşit,dacă $P(c) = 0$,atunci $P_o'(c) \cdot P_1(c) = P'(c) \cdot P'(c) = P'^2(c) > 0$.Prin urmare,şirul construit prin formulele (2.3) este şir Sturm pentru P.Să dovedim că $N(a,b) = V(a) - V(b)$.Fie $x_1,x_2,...,x_r$ rădăcinile reale ale polinoamelor P_i $(i=0,1,..,s-1)$ din intervalul $[a,b]$.Presupunem că $x_1 < x_2 < .. < x_r$.Atunci variaţia $V(x)$ rămîne constantă în fiecare interval (x_k,x_{k+1}).Intr-adevăr,să presupunem că pe un anumit interval (x_k,x_{k+1}) variaţia $V(x)$ nu este constantă.Aceasta înseamnă că există $\alpha,\beta \in (x_k,x_{k+1})$ astfel încît pentru un anumit i avem $P_i(\alpha) \cdot P_i(\beta) < 0$.Cum P_i este continuă,ar exista $c \in (\alpha,\beta)$ astfel încît $P_i(c) = 0$.Deoarece $c \notin \{x_1,x_2,...,x_r\}$ am obţinut o contradicţie.Pentru a termina demonstraţia este sufici-

ent să arătăm că dacă avem un singur y real şi numai unul astfel încît $a<y<b$ şi y este o rădăcină a unui polinom oarecare P_i atunci diferenţa $V(a) - V(b)$ este egală cu 1 cînd y este o rădăcină a lui P şi cu 0 în caz contrar. Să presupunem mai întîi că $P_i(y) = 0$ pentru un anumit i ($1 \le i \le s-1$). Atunci, după 3, $P_{i-1}(y)P_{i+1}(y) < 0$, iar aceste semne nu se schimbă dacă înlocuim pe y prin a sau b. Deci variaţiile de semn în şirurile $P_{i-1}(a), P_i(a), P_{i+1}(a)$ şi $P_{i-1}(b), P_i(b), P_{i+1}(b)$ sînt egale şi anume egale cu 1. Deci dacă y nu este rădăcină a lui P, atunci $V(a)=V(b)$. Dacă y este rădăcină a lui P, $P(a) \cdot P(b) < 0$ iar $P'(a)$ şi $P'(b)$ au acelaşi semn cu $P_1(y)$. Prin urmare $V(a)=V(b)+1$ sau $V(a)-V(b)=1$.

Observaţia 2.2 Demonstraţia ultimei părţi a acestei teoreme se poate folosi şi pentru Teorema 2.3.

Observaţia 2.3 Pentru a obţine numărul tuturor rădăcinilor reale ale lui P aplicăm Teorema 2.4 pe intervalul $(-\infty, +\infty)$.

Exemplul 2.1 Să separăm, folosind un şir Sturm, rădăcinile reale ale polinomului $P(x) = x^3 + x^2 - 2x - 1$. Şirul Strum asociat este:

$P_0 = P$, $P_1 = P' = 3X^2 + 2X - 2$; $P_2 = 2X + 1$, $P_3 = 1$.

Acestui şir i se asociază următorul tablou

	P_0	P_1	P_2	P_3	V
$-\infty$	-	+	-	+	$V(-\infty) = 3$
$+\infty$	+	+	+	+	$V(+\infty) = 0$

Deoarece $V(-\infty) - V(+\infty) = 3$, numărul rădăcinilor reale ale polinomului $P(x)$ este 3. Dacă dorim să găsim intervalele în care se găsesc aceste 3 rădăcini, determinăm numărul de variaţii în punctele 0, ±1, ±2. Rezultatele le prezentăm în următorul tablou:

	P_0	P_1	P_2	P_3	V
-2	-	+	-	+	3
-1	+	-	-	+	2
0	-	-	+	+	1
1	-	+	+	+	1
2	+	+	+	+	0

Deci rădăcinile lui $P(x)$ se găsesc în intervalele $(-2,-1),(-1,0)$ şi respectiv $(1,2)$.

Exemplul 2.2 Să arătăm că pentru funcţia $f(x)=x^2-\sin x-1$, funcţiile $f_0=f$, $f_1(x)=x-\cos x$ şi $f_2(x)=2-\sin x$ formează un şir Sturm pe intervalul $\left[0,\frac{\pi}{2}\right]$, şi să stabilim numărul de rădăcini ale ecuaţiei situate în acest interval.

Condiţiile 1° şi 2° sînt evident satisfăcute. Dacă $c\in\left(0,\frac{\pi}{2}\right)$ are proprietatea $f_1(c)=0$ atunci $f_0(c)\cdot f_2(c)=(c^2-\sin c-1)(2-\sin c)=$ $=(-\sin^2 c-\sin c)(2-\sin c)<0$ (deoarece $c=\cos c$), deci este satisfăcută condiţia 3; dacă $c^2-\sin c-1=0$, atunci $c^2-1=\sin c$ ($c\in\left(0,\frac{\pi}{2}\right)$), deci $c>1$ şi $f'(c)\cdot f_1(c)=(2c-\cos c)(c-\cos c)>0$. Prin urmare, funcţiile date formează un şir Sturm pentru funcţia f. Cum $V(0)=1$ şi $V(\frac{\pi}{2})=0$, rezultă că ecuaţia $f(x)=0$ are o singură rădăcină în $\left[0,\frac{\pi}{2}\right]$.

Incheiem cu observaţia că utilizarea teoremei lui Sturm pentru ecuaţiile transcendente prezintă incovenienţe care ţin de găsirea şirului Sturm asociat.

§3. Metode de determinare a rădăcinilor unei ecuaţii algebrice

Pentru astfel de ecuaţii teorema fundamentală a algebrei ne asigură existenţa unui număr de rădăcini reale sau complexe, eventual multiple, egal cu gradul polinomului. Datorită proprietăţilor polinoamelor există metode specifice de determinare aproximativă a rădăcinilor unei ecuaţii algebrice deosebit de eficiente. Aceste metode sînt de două feluri: 1) metode globale prin care se determină toate rădăcinile ecuaţiei şi 2) metode locale prin care se determină - printr-un procedeu iterativ - cîte o singură soluţie (aproximativă) localizată în prealabil. Deoarece aproximările obţinute prin metode globale au o precizie mică, ele trebuiesc îmbunătăţite cu aproximări iterative obţinute cu o metodă locală. Cea mai cunoscută metodă globală este metoda Lobacevski-Graeffe pe care o prezentăm în cele ce urmează.

Metoda Lobacevschi-Graeffe

Considerăm ecuaţia algebrică

(3.1) $$P(x) = x^n + a_1 x^{n-1} + \dots + a_{n-1}x + a_n = 0$$

cu rădăcinile $x_1, x_2, \dots, x_n$. Vom construi şirul polinoamelor $P_k(x)$, ale căror rădăcini sînt $-x_1^{2^k}, \dots, -x_n^{2^k}$, de forma:

$$P_k(x) = x^n + a_1^{(k)}x^{n-1} + \dots + a_{n-1}^{(k)}x + a_n^{(k)} .$$

Pentru a deduce o relaţie de recurenţă care să permită determinarea coeficienţilor $a_j^{(k)}$ observăm mai întîi că dacă $y_1, \dots, y_n$ sînt rădăcinile lui P_k, atunci P_{k+1} are rădăcinile $-y_1^2, \dots, -y_n^2$. Intr-adevăr, dacă $y_i = -x_i^{2^k}$, atunci $-x_i^{2^{k+1}} = -(-x_i^{2^k})(-x_i^{2^k}) = -y_i^2$ sînt rădăcinile lui $P_{k+1}(x)$. Au loc relaţiile

$$P_k(x) = (x-y_1)(x-y_2)\dots(x-y_n), \qquad P_{k+1}(x) = (x+y_1^2)\dots(x+y_n^2).$$

Rezultă că $P_{k+1}(-x^2) = (-x^2+y_1^2)\dots(-x^2+y_n^2) = P_k(-x)\cdot P_k(x)$, adică au loc egalităţile

$$\sum_{i=0}^{n} a_i^{(k+1)}(-x^2)^{n-i} = \Big(\sum_{j=0}^{n} a_j^{(k)}(-x)^{n-j}\Big)\Big(\sum_{s=0}^{n} a_s^{(k)}x^{n-s}\Big)$$

şi

$$\sum_{i=0}^{n} (-1)^{n-i} a_i^{(k+1)} x^{2n-2i} = \sum_{j,s=0}^{n} (-1)^{n-j} a_j^{(k)} a_s^{(k)} x^{2n-(j+s)} .$$

Identificînd coeficienţii celor două polinoame se obţine

$$(-1)^{n-i} a_i^{(k+1)} = \sum_{j+s=2i} (-1)^{n-j} a_j^{(k)} a_s^{(k)} .$$

Făcînd notaţia $j-i=m$ rezultă $s=i-m$ şi în plus

$$a_i^{(k+1)} = \sum_{m=-p}^{p} (-1)^m a_{i+m}^{(k)} a_{i-m}^{(k)} \qquad p = \min\{i, n-i\} ,$$

sau

(3.2) $$a_i^{(k+1)} = (a_i^{(k)})^2 + 2\sum_{m=1}^{p} (-1)^m a_{i+m}^{(k)} a_{i-m}^{(k)} .$$

Fără demonstraţie prezentăm următoarele două rezultate:

Propoziţia 3.1 Dacă polinomul cu coeficienţi reali $P(x)$ are rădăcinile $x_1, x_2, \dots, x_n$ reale şi dacă aceste rădăcini sînt numerotate astfel încît $|x_1| \geq |x_2| \geq \dots \geq |x_n|$, atunci

(3.3) $$|x_j| = \lim_{k\to\infty} \sqrt[2^k]{\frac{a_j^{(k)}}{a_{j-1}^{(k)}}} \qquad j \in \{1, 2, \dots, n\}$$

şi în plus $a_j^{(k)} \geq 0$ pentru orice j şi k.

Propoziţia 3.2 Dacă P este un polinom cu coeficienţi reali ale cărui rădăcini $x_1, \dots, x_n$ verifică relaţiile

$$|x_1| = \dots = |x_{m_1}| > |x_{m_1+1}| = \dots = |x_{m_2}| > \dots$$

atunci

(3.6) $$|x_{m_s}|^{m_s - m_{s-1}} = \lim_{k\to\infty} \sqrt[2^k]{\frac{a^{(k)}_{m_s}}{a^{(k)}_{m_{s-1}}}} .$$

Pe baza Propoziţiilor 3.1 şi 3.2 se stabilesc următorii algoritmi de determinare a rădăcinilor ecuaţiei (3.1).

Dacă

(3.7) $$|x_1| > |x_2| > \dots > |x_n|$$

algoritmul este:

1) se calculează coeficienţii polinoamelor P_k cu formulele (3.2),

2) pentru fiecare polinom se cercetează dacă există coeficienţi negativi; dacă există coeficienţi negativi, înseamnă că nu este satisfăcută condiţia (3.7), adică (3.1) are rădăcini complex conjugate

3) se verifică dacă

(3.8) $$\left| \frac{a_i^{(k+1)}}{(a_i^{(k)})^2} - 1 \right| \leq \varepsilon \qquad \forall\ i=1,2,\dots,n-1$$

cu ε convenabil ales, apropiat de eroarea minimă admisă de calculator,

4) dacă este satisfăcută condiţia 3), se calculează modulele rădăcinilor cu formula

$$|x_i| = \sqrt[2^k]{a_i^{(k)}/(a_{i-1}^{(k)})^2} \qquad i=1,2,\dots,n.$$

In caz contrar, se reiau calculele cu un nou polinom din şir şi acest ciclu se repetă de un număr de N ori. Dacă inegalităţile (3.8) nu sînt satisfăcute după acest număr de cicli, se consideră că (3.7) nu este adevărată.

5). Inlocuind în ecuaţie $\pm|x_i|$, determinăm semnul rădăcinii alegînd cel mai mic dintre numerele $|f(|x_i|)|$ şi $|f(-|x_i|)|$. Valoarea minimă obţinută este în acelaşi timp o măsură a preciziei cu care se calculează rădăcina respectivă. *

Dacă

(3.9) $$|x_1| = |x_2| = \dots = |x_{m_1}| > |x_{m_1+1}| = \dots = |x_{m_2}| > \dots > |x_{m_{p-1}+1}| = \dots = |x_n|$$

algoritmul este următorul:

1) se calculează coeficienţii polinoamelor P_k pentru $k=1,2,..,N$ cu formulele de recurenţă (3.2)

2) pentru $k=N-1$ se cercetează pentru ce indici $i \in \{1,2,..,n-1\}$ este satisfăcută inegalitatea (3.8). Adăugînd indicii 0 şi n pentru care inegalitatea este automat satisfăcută, obţinem $m_0 = 0 < m_1 < m_2 < \ldots < m_p = n$ din relaţiile (3.9). Multiplicităţile modulelor rădăcinilor sînt atunci $m_i - m_{i-1}$ $(i=1,2,..,p)$

3) se aproximează modulele rădăcinilor cu formulele

$$\left|x_{m_i}\right|^{m_i-m_{i-1}} = \sqrt[2^k]{a_{m_i}^{(N)}/a_{m_{i-1}}^{(N)}} \qquad i=1,2,..,p$$

4) se încearcă cu schema lui Horner, pe rînd, dacă $\pm\left|x_{m_i}\right|$ sînt rădăcini şi, în caz afirmativ, se determină multiplicităţile lor; înlăturînd aceste rădăcini prin împărţire, rămîne un polinom care are numai rădăcini complex conjugate de module cunoscute şi multiplicităţi cunoscute.

Determinarea rădăcinilor complex conjugate se face în felul următor. Fie $x=x_{m_i}$ o rădăcină complexă al cărei modul a fost calculat ca mai sus. Cum conjugata lui x este tot o rădăcină, polinomul P(x) se divide cu un trinom de forma x^2+px+q în care $q = |x|^2$ este cunoscut. Rămîne să determinăm parametrul p astfel ca restul împărţirii la x^2+px+q să fie nul. Scriind identitatea:

$$x^n+a_1x^{n-1}+..+a_{n-1}x+a_n=(x^2+px+q)(x^{n-2}+b_1x^{n-3}+..+b_{n-2})+b_{n-1}x+b_n$$

obţinem prin identificare

$$b_1=a_1-p\ ,\ b_2=a_2-pb_1-q\ ,\ldots,\ b_{n-2}=a_{n-2}-pb_{n-3}-qb_n,$$

$$b_{n-1}=a_{n-1}-pb_{n-2}-qb_{n-3}\ ,\quad b_n=a_n-qb_{n-2}\ .$$

Dacă notăm $b_{-1}=0$, $b_0=1$, $a_0=1$ obţinem relaţiile de recurenţă

$$(3.10)\qquad \begin{cases} b_k = a_k - pb_{k-1} - qb_{k-2} & k=1,2,..,n-1 \\ b_n = a_n - qb_{n-2} & . \end{cases}$$

Condiţia de divizibilitate este:

$$(3.11)\qquad b_{n-1} = a_{n-1} - pb_{n-2} - qb_{n-3} = 0$$

(3.11$_2$) $$b_n = a_n - qb_{n-2} = 0 .$$

Formulele (3.10) ne arată că b_{n-1} este un polinom de gradul n-1 şi b_n - un polinom de gradul n în p. Pentru a găsi parametrul p ar trebui determinate rădăcinile comune celor două polinoame. Ştiind că orice rădăcină comună polinoamelor b_n şi b_{n-1} este rădăcină şi pentru combinaţia lor liniară

$$b_n^* = b_{n-1} - \frac{p}{q} b_n = a_{n-1} - \frac{p}{q} a_n - qb_{n-3} ,$$

care este un polinom de gradul 3; calculăm rădăcinile reale ale lui b_n^* cu algoritmul de mai sus. Incercînd pentru fiecare din ele dacă b_{n-1} şi b_n sînt destul de mici în valoare absolută, se obţin toate rădăcinile de modul dat ale lui P(x).

Exemplul 3.1 (Demidovici) Să determinăm prin metoda Lobacevski-Graeffe rădăcinile ecuaţiei $x^3 - 3x + 1 = 0$. Fie $P(x) = x^3 - 3x + 1$ şi $P_1(-x^2) = P(-x)P(x)$. Se găseşte că $P_1(-x^2) = -x^6 + 6x^4 - 9x^2 + 1$, adică $P_1(x) = x^3 + 6x^2 + 9x + 1$. Coeficienţii lui P_1 se puteau determina şi cu ajutorul formulei (3.2). Rezultatele calculelor, cu 4 cifre semnificative sînt conţinute în următorul tabel

Puteri	x^3	x^2	x	x^0
1	1	0	-3	1
2	1	6	9	1
4	1	18	69	1
8	1	$1,86\cdot10^2$	$4725\cdot10^3$	1
16	1	$2,515\cdot10^4$	$2,233\cdot10^7$	1
32	1	$5,878\cdot10^8$	$4,986\cdot10^{14}$	1
64	1	$3,455\cdot10^{17}$	$2,486\cdot10^{29}$	1

Rezultă că

$$x_1 = \pm\sqrt[64]{3,445\cdot10^{17}} = \pm1,879 \qquad x_2 = \pm\sqrt[64]{\frac{2,486}{3,445}\cdot10^{12}} = \pm1,532$$

$$x_3 = \pm\sqrt[64]{\frac{1}{2,486}\cdot10^{-29}} = \pm0,347$$

Deoarece $x_1 + x_2 + x_3 = 0$, rezultă că cel puţin o rădăcină este negativă. Cum $P(+\infty) = +\infty$ şi $P(0) = 1$, rezultă că polinomul are două rădăcini pozitive şi una negativă. Totodată rezultă că rădăcina negativă are valoarea absolută mai mare şi prin urmare

$$x_1 = -1,879 \qquad x_2 = 1,532 \qquad x_3 = 0,347 .$$

§4. Metode iterative de rezolvare a ecuației $f(x)=0$

Metodele iterative de rezolvare numerică a ecuației

$$f(x) = 0 \tag{4.1}$$

pe care urmează să le prezentăm, fac parte din clasa metodelor locale. Așa cum am precizat la începutul §3, metodele locale se aplică pentru aproximarea unei soluții ξ localizate într-un interval $[a,b]$. (Înseamnă că $f(\xi)=0$ și că $f(a)\cdot f(b)<0$).

Metoda bisecției

Presupunem că în intervalul $[a,b]$ se află o singură rădăcină ξ a ecuației (4.1). Deci $f(\xi)=0$ și $f(a)\cdot f(b)<0$. Se împarte intervalul $[a,b]$ în două părți egale, după care se verifică în care jumătate se află soluția ξ cu ajutorul semnului lui $f(\frac{a+b}{2})$ (dacă $f(\frac{a+b}{2})\neq 0$). Notăm cu $[a_1,b_1]$ noul interval în care se află soluția. Acest segment se împarte și el în părți egale și se determină un nou segment $[a_2,b_2]$ care conține rădăcina. Repetînd acest procedeu se obține un șir de segmente incluse unul în celălalt $[a_1,b_1]$, $[a_2,b_2]$, ..., $[a_n,b_n]$, ... astfel încît $f(a_n)\cdot f(b_n)<0$ și cu $b_n-a_n = \frac{1}{2^n}(b-a)$. Extremitățile $a_1,a_2,\ldots,a_n,\ldots$ și $b_1,b_2,\ldots,b_n,\ldots$ formează două șiruri, unul crescător iar celălalt descrescător, care vor avea aceeași limită ξ^*. Ținînd cont de continuitatea lui f rezultă că $\lim\limits_{n\to\infty} f(a_n)\cdot f(b_n) = [f(\xi^*)]^2 \leq 0$, adică $f(\xi^*)=0$. Cum în intervalul $[a,b]$ este o singură soluție a ecuației (4.1), rezultă că $\xi^*=\xi$. Dacă aproximăm soluția ξ prin a_n sau b_n, eroarea comisă este $<\frac{1}{2^n}(b-a)$.

Metoda coardei

Principiul acestei metode este următorul: dacă soluția x^* a fost localizată ca fiind în intervalul $[a,b]$, aproximăm x^* cu $\tilde{x}\in(a,b)$ definit de proprietatea ca $\tilde{x}-a$ și $-b+\tilde{x}$ să fie proporționale cu $f(a)$ și $f(b)$. Se poate dovedi următoarea:

Teorema 4.1 Fie $f:[a,b]\longrightarrow R$, o funcție de două ori derivabilă, care are proprietățile $f'(x)\neq 0$, $f''(x)\neq 0$, $\forall x\in[a,b]$ și $f(a)\cdot f(b)<0$.

Atunci:

a) ecuaţia $f(x)=0$ are o soluţie unică $x^* \in (a,b)$ care se obţine ca limită a şirului strict monoton definit prin

(4.2) $x_0=a,\ x_n=x_{n-1}-\dfrac{f(x_{n-1})}{f(x_{n-1})-f(b)}(x_{n-1}-b),\ n\geq 1,$ dacă $f(a)f''(a)<0$

şi

(4.3) $x_0=b,\ x_n=x_{n-1}-\dfrac{f(x_{n-1})}{f(x_{n-1})-f(a)}(x_{n-1}-a),\ n\geq 1,$ dacă $f(b)f''(b)<0.$

b) Dacă în plus există $m_1, M_1>0$ astfel încît $m_1<|f'(x)|<M_1, \forall x\in [a,b]$, atunci x^* satisface inegalităţile

(4.4) $$|x^*-x_n|\leq \frac{1}{m_1}|f(x_n)|$$

(4.5) $$|x^*-x_n|\leq\frac{M_1-m_1}{m_1}|x_n-x_{n-1}| \qquad n\geq 1.$$

Demonstraţie f fiind continuă şi $f(a)\cdot f(b)<0$, rezultă că f are cel puţin o soluţie. Cum $f'(x)\neq 0, \forall x\in[a,b]$ şi f' are proprietatea lui Darboux, avem $f'>0$ sau $f'<0$ pe $[a,b]$, deci f este strict monotonă şi în consecinţă (4.1) are exact o soluţie, pe care o vom nota cu x^*. Presupunem că $f(a)\cdot f''(a)<0$ şi $f(a)<0$. Deoarece $f''(x)\neq 0, \forall x\in[a,b]$, rezultă că $f''>0$ pe $[a,b]$, adică f este convexă. Ecuaţia dreptei care trece prin punctele $P(b,f(b))$ şi $P_n(x_n,f(x_n))$ este

$$y=f(x_n)+\frac{f(x_n)-f(b)}{x_n-b}(x-x_n) .$$

Această dreaptă intersectează axa Ox în punctul de abscisă

$$\tilde{x} = x_n-\frac{x_n-b}{f(x_n)-f(b)}f(x_n)$$

adică $x_{n+1}=\tilde{x}$ este abscisa punctului obţinut prin intersecţia coardei P_nP cu axa Ox. Deci s-a obţinut interpretarea geometrică a metodei (care justifică şi denumirea metodei). Ţinînd cont de convexitatea funcţiei f pe intervalul $[a,b]$, dacă $a\leq x_n\leq x^*$ va rezulta că $f(x_{n+1})<0=f(x^*)$ (graficul funcţiei se află sub graficul coardei) şi că $x_{n+1}\leq x^*$. Deoarece $f(x_n)<0$ şi $f'(x_n)>0$ şi ţinînd cont de teorema lui Lagrange, rezultă că există $c_n\in(x_n,b)$ astfel încît au loc relaţiile:

$$x_{n+1}=x_n-\frac{x_n-b}{f(x_n)-f(b)}f(x_n)=x_n-\frac{f(x_n)}{f'(c_n)}>x_n .$$

Am dovedit deci că şirul $\{x_n\}_{n\in N}$ este strict crescător şi mărginit, adică este convergent. Dacă $\tilde{x} = \lim_{n\to\infty} x_n$, atunci $\tilde{x} \le x^*$ şi, trecînd la limită în relaţia (4.2) se obţine

$$\tilde{x} = \tilde{x} - \frac{f(\tilde{x})}{f(\tilde{x}) - f(b)}(\tilde{x} - b),$$

adică $f(\tilde{x}) = 0$. In virtutea unicităţii soluţiei rezultă $\tilde{x} = x^*$.

b) Fie $x_n \in [a,b]$ fixat. Conform teoremei lui Lagrange există $c \in (x_n, x^*)$ astfel încît

$$f(x^*) - f(x_n) = f'(c)(x^* - x_n)$$

adică

$$|x^* - x_n| = \frac{|f(x_n)|}{|f'(c)|} \le \frac{|f(x_n)|}{m_1} \qquad \forall n \ge 1.$$

Din formula (4.2) pusă sub forma

$$f(x^*) - f(x_{n-1}) = -f(x_{n-1}) = \frac{f(x_{n-1}) - f(b)}{x_{n-1} - b}(x_n - x_{n-1})$$

$(n \ge 1)$ şi din teorema lui Lagrange rezultă că există $\xi_n \in (x_n, x^*)$ şi $\eta_n \in (x_{n-1}, b)$ astfel încît

$$(x^* - x_{n-1})f'(\xi_n) = (x_n - x_{n-1})f'(\eta_n)$$

sau

$$x^* - x_{n-1} = \frac{f'(\eta_n)}{f'(\xi_n)}(x_n - x_{n-1}).$$

Au loc egalităţile

$$x^* - x_n = (x^* - x_{n-1}) - (x_n - x_{n-1}) = \left(\frac{f'(\eta_n)}{f'(\xi_n)} - 1\right)(x_n - x_{n-1})$$

din care, ţinîndu-se cont că f' are semn constant, se deduce

$$|x^* - x_n| = \frac{|f'(\eta_n) - f'(\xi_n)|}{|f'(\xi_n)|}|x_n - x_{n-1}| \le \frac{M_1 - m_1}{m_1}|x_n - x_{n-1}|.$$

Incheiem cu observaţia că $f(a)\cdot f''(a) < 0$ şi $f(b)\cdot f''(b) < 0$ fiind complementare una faţă de alta teorema cuprinde toate cazurile posibile.

Metoda tangentei (Newton-Raphson)

Baza acestei metode este următorul rezultat

Teorema 4.2 Fie $f:[a,b] \longrightarrow R$ o aplicaţie de două ori derivabilă, cu $f'(x) \ne 0$, $f''(x) \ne 0$, $\forall x \in [a,b]$ şi $f(a)\cdot f(b) < 0$. Atunci:

a) ecuaţia $f(x) = 0$ are soluţie unică $x^* \in (a,b)$ care poate fi obţinută ca limită a şirului $\{x_n\}_{n\in N}$ definit astfel

(4.6) $$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \qquad \forall n \ge 1$$

unde $x_0 \in [a,b]$ este fixat arbitrar astfel încît $f(x_0)f''(x_0) > 0$.

b) Au loc inegalitățile

(4.7) $$|x^* - x_n| \leq \frac{1}{m_1}|f(x_n)| \qquad \forall n \geq 1,$$

(4.8) $$|x^* - x_n| \leq \frac{M_2}{2m_1}(x_n - x_{n-1})^2 \qquad \forall n \geq 1,$$

(4.9) $$|x^* - x_n| \leq \frac{M_2}{2m_1}(x^* - x_{n-1})^2 \qquad \forall n \geq 1,$$

unde $m_1 = \inf_{x\in[a,b]} |f'(x)|$ și $M_2 = \sup_{x\in[a,b]} |f''(x)|$.*

Demonstrație a) Existența soluției unice x^* se stabilește la fel ca în demonstrația teoremei 4.1. Să presupunem că $f(a)<0$ și deci $f(b)>0$. Atunci f, fiind strict monotonă, va fi strict crescătoare, adică $f'>0, \forall x\in[a,b]$. Din condiția $f(x_0)f''(x_0)>0$ rezultă sau $f''(x_0)>0$ și $f(x_0)>0$ sau $f''(x_0)<0$ și $f(x_0)<0$. Să presupunem că sîntem în primul caz, deci avem $x_0 > x^*$; să arătăm, prin inducție completă, că $x_n > x^*, \forall n \geq 0$. Să presupunem că $x_n > x^*$. Folosind formula lui Taylor cu restul lui Lagrange, obținem

(4.10) $$0 = f(x^*) = f(x_n) + \frac{x^* - x_n}{1!}f'(x_n) + \frac{(x^* - x_n)^2}{2!}f''(\xi_n)$$

unde ξ_n este cuprins între x^* și x_n. Deoarece $f''(\xi_n)>0$, din (4.10) rezultă $f(x_n) + (x^*-x_n)f'(x_n) < 0$, adică

(4.11) $$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > x^* .$$

Conform principiului inducției complete rezultă că $x_n > x^*, \forall n \geq 0$. Funcția f fiinnd strict crescătoare, rezultă că $f(x_n) > f(x^*)=0$ $(\forall n \geq 0)$ și $f'(x_n)>0$, iar relația (4.11) ne conduce la inegalitatea $x_{n+1} < x_n$, adică șirul $\{x_n\}_{n\in N}$ este strict descrescător. Cum $\{x_n\}_{n\in N}$ este minorat de a, rezultă că este convergent. Dacă $\tilde{x} = \lim_{n\to\infty} x_n$, trecînd la limită în relația (4.6) rezultă că $\tilde{x} = \tilde{x} - \frac{f(\tilde{x})}{f'(\tilde{x})}$, adică $f(\tilde{x})=0$. Din unicitatea soluției ecuației (4.1) pe (a,b), rezultă că $\tilde{x} = x^*$.

b) Fie $x \in [a,b]$ fixat. Conform teoremei lui Lagrange există c cuprins între x și x^* astfel încît

$$f(x^*) - f(x) = (x^* - x)f'(c)$$

şi prin urmare are loc

$$|x^* - x| = \frac{|f(x)|}{|f'(c)|} \leq \frac{|f(x)|}{m_1} \qquad \forall n \geq 1 .$$

Aplicînd formula lui Taylor se obţine

(4.12) $$f(x_n)=f(x_{n-1})+\frac{x_n-x_{n-1}}{1!} f'(x_{n-1}) + \frac{(x_n-x_{n-1})^2}{2!} f''(c_n)$$

unde c_n este cuprins între x_{n-1} şi x_n.Din (4.6) rezultă

(4.13) $$f(x_{n-1}) + (x_n - x_{n-1})f'(x_{n-1}) = 0 .$$

Combinînd (4.12) cu (4.13) deducem

$$f(x_n) = \frac{(x_n-x_{n-1})^2}{2} f''(c_n) ,$$

de unde,ţinînd cont de (4.7) obţinem

$$|x^*-x_n| \leq \frac{1}{2m_1} (x_n-x_{n-1})^2 |f''(c_n)| \leq \frac{M_2}{2m_1} (x_n-x_{n-1})^2 .$$

Din (4.10) rezultă

(4.14) $$x^* - x_n = -\tfrac{1}{2}(x^* - x_n)^2 \frac{f''(\xi_n)}{f'(x_n)} - \frac{f(x_n)}{f'(x_n)} .$$

Relaţiile (4.6) şi (4.14) conduc la egalitatea

$$x^* - x_{n+1} = -\tfrac{1}{2}(x^* - x_n)^2 \frac{f''(\xi_n)}{f'(x_n)} ,$$

care ne conduce,la rîndul ei,la inegalitatea

$$|x^* - x_{n+1}| \leq \frac{M_2}{2m_1} (x^* - x_n)^2 \quad .*$$

c) <u>Metode iterative bazate pe principiul contracţiei</u>

Teorema de punct fix a lui Banach ne dă un procedeu iterativ de determinare a punctului fix al unei contracţii definite pe un spaţiu metric complet.De obicei ne interesează aplicaţii care nu sînt definite pe întreg spaţiul (sau chiar dacă sînt definite pe întreg spaţiul ne interesează comportarea lor locală).De aceea ne va fi util următorul rezultat.

<u>Propoziţia 4.3</u> <u>Fie</u> (E, ρ) <u>un spaţiu metric complet</u>, $S(x_o,r) = \{y ; y \in E , \rho(x,y) \leq r\}$ <u>o sferă închisă din</u> E <u>şi</u> $T:S(x_o,r) \longrightarrow E$ <u>o aplicaţie care are proprietăţile:</u>

1) <u>există un număr real</u> $\alpha<1$ <u>astfel ca</u> $\rho(Tx,Ty) < \alpha\rho(x,y)$ <u>pentru orice</u> $x,y \in S(x_o,r)$

2) $\rho(x_0, Tx_0) \leq r(1-\alpha)$.

<u>Atunci</u> $T(S(x_0,r)) \subset S(x_0,r)$ <u>şi în plus</u>

a) T <u>este continuă</u>,

b) <u>oricare ar fi</u> $x_0 \in E$, <u>şirul</u> $\{x_n\}_{n\in N}$ <u>definit prin relaţia</u> $x_n = T(x_{n-1})$ <u>este convergent iar limita sa</u> $x^* (= \lim_{n\to\infty} x_n)$ <u>este punct fix pentru</u> T (<u>adică</u> $Tx^* = x^*$) <u>şi</u>

$$\rho(x_n, x^*) \leq \frac{\alpha^n}{1-\alpha} \cdot \rho(x_0, x_1) \ ,$$

c) <u>aplicaţia</u> T <u>are un punct fix unic</u>.

<u>Demonstraţie</u> Pentru un $x \in S(x_0,r)$, oarecare, au loc inegalităţile $\rho(Tx,x_0) \leq \rho(Tx,Tx_0) + \rho(Tx_0,x_0) \leq \alpha\rho(x,x_0) + r(1-\alpha) \leq r$, adică $Tx \in S(x_0,r)$. Prin urmare T poate fi considerată ca fiind o contracţie pe spaţiul metric complet $S(x_0,r)$ şi prin urmare îi putem aplica concluziile teoremei de punct fix a lui Banach (vezi Cap.I), care ne asigură că concluziile a), b) şi c) ale propoziţiei sînt adevărate. □

Să presupunem că ecuaţia (4.1) a fost transformată, printr-un procedeu oarecare, în ecuaţia (local) echivalentă

$$\varphi(x) = x \ . \tag{4.15}$$

Dacă pe domeniul de definiţie sau pe o submulţime închisă (cu interior nevid) a sa φ este o contracţie, putem construi şirul $\{x_n\}_{n\in N}$, pornind de la un element x_0 fixat arbitrar în domeniul pe care φ este contracţie, cu ajutorul relaţiei de recurenţă $x_n = \varphi(x_{n-1})$, care va fi convergent la un element x^*, element care va fi evident şi soluţia ecuaţiei $f(x) = 0$ situată în zona de interes. Problema transformării ecuaţiei (4.1) într-o ecuaţie (4.15) echivalentă cu ea (şi pentru care funcţia φ să fie o contracţie) este o problemă destul de dificilă. Trebuie remarcat mai întîi că echivalenţa se realizează în general local. Intr-adevăr, dacă ecuaţia (4.1) are mai multe rădăcini distincte $x_1, x_2, \dots x_n$ situate fiecare în cîte un interval $x_k \in [a_k, b_k]$ $(k=1,2,\dots,n)$ ne punem problema transformării ei într-o ecuaţie de tip (4.15)

separat pe fiecare interval $[a_k, b_k]$ $(k=1,2,\ldots,n)$. O echivalenţă globală nu este posibilă deoarece ar veni în contradicţie cu unicitatea punctului fix al unei contracţii. In ipoteza că ecuaţia (4.1) a fost pusă sub forma echivalentă (4.15) pe intervalul $[a,b]$, apare problema să recunoaştem dacă φ este sau nu contracţie pe $[a,b]$. Un astfel de criteriu este dat de propoziţia care urmează.

Propoziţia 4.4 Fie $\varphi:[a,b]\longrightarrow R$ o funcţie continuă, derivabilă pe (a,b), astfel ca $\sup\limits_{x\in(a,b)}|\varphi'(x)|<1$. Să presupunem că

$$\left|\varphi\left(\frac{a+b}{2}\right)-\frac{a+b}{2}\right|\le\frac{b-a}{2}(1-\alpha)$$

unde $\alpha=\sup\limits_{x\in(a,b)}|\varphi'(x)|$. Atunci ecuaţia $\varphi(x)=x$ are o rădăcină unică $x^*\in[a,b]$. Dacă $x_1\in[a,b]$ este un element fixat, şirul definit prin relaţia de recurenţă $x_{n+1}=\varphi(x_n)$ va avea limita x^* şi în plus $|x_n-x^*|\le\frac{\alpha^{n-1}}{1-\alpha}|x_2-x_1|$.

Demonstraţie Propoziţia este o consecinţă directă a Propoziţiei 4.3. Intr-adevăr, dacă facem particularizările $E=R$, $x_0=\frac{a+b}{2}$ şi $T=\varphi$ şi ţinem cont de relaţiile

$$|\varphi(x)-\varphi(y)|=|x-y||\varphi'(\xi)|\le\alpha|x-y|\qquad\forall x,y\in[a,b],$$

vor fi satisfăcute ipotezele Propoziţiei 4.3.

Observaţia 4.1 Pentru şirul construit cu teorema de punct fix a lui Banach convergenţa este aceeaşi cu convergenţa seriei geometrice de raţie α. Apare problema reprezentării rădăcinii unei ecuaţii ca limită a unui şir convergent mai repede decît o progresie geometrică. O astfel de posibilitate este evidenţiată în propoziţia care urmează.

Propoziţia 4.5 Fie $\varphi:[a,b]\longrightarrow R$ o funcţie astfel încît:

a) φ este de p ori derivabilă pe $[a,b]$, $p\ge2$ şi $\varphi^{(p)}$ este mărginită pe $[a,b]$,

b) ecuaţia $x=\varphi(x)$ are rădăcina $x^*\in(a,b)$,

c) $\varphi'(x^*)=\ldots=\varphi^{(p-1)}(x^*)=0$.

Fie $M = \frac{1}{p!}\sup_{x\in[a,b]}|\varphi^{(p)}(x)|$ şi x_0 astfel încît $|x_0-x^*|\le\min(1,x^*-a,b-x^*)$ şi $M|x_0-x^*|\le\alpha<1$. Atunci se poate forma şirul $\{x_n\}_{n\in N}$ prin relaţia de recurenţă $x_{n+1}=\varphi(x_n)$, convergent la x^*, ale cărui elemente verifică relaţia

(4.16) $$|x_n - x^*|\le \alpha^{\frac{p^n-1}{p-1}} .$$

Demonstraţie. Vom arăta inductiv că, în ipotezele propoziţiei noastre, şirul $\{x_n\}_{n\in N}$ se poate forma. Fie $\delta=\min(1,x^*-a,b-x^*)$. Atunci $|x_0-x^*|\le\delta$. Presupunem că $|x_n-x^*|\le\delta$ şi că $M|x_n-x^*|\le\alpha$. Ţinînd cont că φ este de p ori derivabilă pe (a,b) se deduc egalităţile

$$|x_{n+1}-x^*| = |\varphi(x_n)-\varphi(x^*)| = \Big|(x_n-x^*)\varphi'(x^*)+\dots+\frac{(x_n-x^*)^{p-1}}{(p-1)!}\varphi^{(p-1)}(x^*)+ \frac{(x_n-x^*)^p}{p!}\varphi^{(p)}(\xi)\Big|,$$

care, împreună cu condiţia c), conduc la

(4.17) $$|x_{n+1}-x^*| = \frac{1}{p!}|\varphi^{(p)}(\xi)||x_n-x^*|^p \le M\cdot|x_n-x^*|^p .$$

Din ipoteza inductivă şi din relaţia (4.17) rezultă că $|x_{n+1}-x^*|\le\delta$ şi că $M|x_{n+1}-x^*|\le M^2|x_n-x^*|^p\le\alpha$. Conform principiului inducţiei complete, elementele şirului $\{x_n\}_{n\in N}$ sînt conţinute în (a,b). In plus au loc relaţiile

$$|x_n-x^*|\le M[M[\dots[M[x_0-x^*|^p]^p\dots]^p = M^{1+p+\dots+p^{n-1}}|x_0-x^*|^{p^n} = M^{\frac{p^n-1}{p-1}}|x_0-x^*|^{p^n} \qquad (p\ge2).$$

Deci are loc inegalitatea

$$|x_n-x^*|\le (M|x_0-x^*|)^{\frac{p^n-1}{p-1}}|x_0-x^*|^{p^n-\frac{p^n-1}{p-1}} \qquad (p\ge 2).$$

Inegalitatea precedentă şi condiţia $|x_0-x^*|\le\delta$, implică inegalitatea (4.16).

Rămîne de rezolvat problema : avînd o funcţie $f:[a,b]\longrightarrow R$ astfel încît ecuaţia $f(x)=0$ să aibă o singură soluţie $x^*\in(a,b)$, să se construiască o funcţie φ, definită pe un interval $(c,d)\subset(a,b)$ care să conţină pe x^*, care să satisfacă condiţiile a), b),

şi c) ale propoziţiei 4.5.

Răspunsul la această problemă este conţinut în următoarea propoziţie, pe care o prezentăm fără demonstraţie.

Propoziţia 4.6 Fie f o funcţie continuă pe $[a,b]$, derivabilă de 2p-1 ori ($p \geq 2$) cu $f^{(2p-1)}$ mărginită pe $[a,b]$ şi $x^* \in (a,b)$, astfel încît $f(x^*)=0, f'(x^*) \neq 0$. Atunci:

a) există un interval $[c,d] \subset [a,b]$ astfel încît $x^* \in (c,d), f'(x) \neq 0$ pe $[c,d]$ (deci f strict monotonă pe $[c,d]$),

b) funcţiile $b_1(x)=\frac{1}{f'(x)}, b_2(x), \ldots, b_{p-1}(x)$ definite prin relaţia de recurenţă

$$b_{k+1}(x) = \frac{b_k'(x)}{f'(x)}$$

sînt de p ori derivabile pe $[c,d]$ şi $b_1^{(p)}, \ldots, b_{p-1}^{(p)}$ sînt mărginite pe $[c,d]$.

c) funcţia $\varphi: [c,d] \longrightarrow R$ definită prin

$$\varphi(x) = x - f(x)\frac{b_1(x)}{1!} + \ldots + (-1)^{p-1}[f(x)]^{p-1} \frac{b_{p-1}(x)}{(p-1)!}$$

satisface condiţiile $\varphi(x^*)=x^*, \varphi'(x^*)=\ldots=\varphi^{(p-1)}(x^*)=0$ şi $\varphi^{(p)}$ există şi este mărginită pe $[c,d]$. *

Remarcăm faptul că în cazul p=2, $\varphi(x)=x-\frac{f(x)}{f'(x)}$, iar procedeul iterativ corespunzător este metoda Newton-Raphson. De asemenea trebuie remarcat faptul că dacă ecuaţia (4.1) este o ecuaţie algebrică, procedeul indicat de propoziţia 4.5 se poate modifica în aşa fel încît funcţia φ să fie polinom.

§5. Exerciţii

1) Să se formeze şirul Sturm şi să se separe rădăcinile polinoamelor:

a) x^3-3x+5, b) $2x^4-8x^3+8x^2-1$, c) $x^4-2x^3-3x^2+2x+1$,
d) x^3-7x+7, e) $3x^4+12x^3+9x^2-1$, f) $9x^4-126x^2-252x-140$.

2) Să se arate că funcţiile $f(x)=3x^2-e^x$, $f_1(x)=5x-e^x$, $f_2(x)=6-e^x$ formează un şir Sturm pe intervalul $[0,1]$ pentru

ecuaţia $3x^2 - e^x = 0$ şi să se determine numărul de rădăcini ale acestei ecuaţii situate în acest interval.

3) Determinaţi rădăcina pozitivă a ecuaţiei

$$x^3 - 0,2\,x^2 - 0,2\,x - 1,2 = 0$$

cu o eroare $< 2\cdot 10^{-3}$, folosind metoda coardei.

4) Calculaţi prin metoda Newton-Raphson rădăcina negativă a ecuaţiei

$$x^4 - 3x^2 + 75x - 10000 = 0$$

cu cinci cifre semnificative exacte.

5) Determinaţi prin metoda Newton-Raphson rădăcina pozitivă cea mai mică a ecuaţiei $\mathrm{tg}\,x = x$, cu o eroare $< 10^{-4}$.

6) Determinaţi cu o eroare $< 5\cdot 10^{-4}$ rădăcina pozitivă a ecuaţiei

$$x^5 - x - 0,2 = 0.$$

7) Calculaţi rădăcinile reale ale ecuaţiei $x = \sin x + 0,25$ cu trei cifre semnificative exacte, folosind metoda aproximaţiilor succesive.

8) Să se determine soluţia ecuaţiei $1 - \ln x = 0$ situată în intervalul $[2,3]$, cu o eroare $< 10^{-3}$, prin metodele:

a) metoda bisecţiei

b) metoda coardei

c) metoda Newton-Raphson.

9) Folosind metoda Lobacevski-Graeffe calculaţi rădăcinile ecuaţiilor:

a) $x^3 - 3x + 1 = 0$,

b) $x^4 + x^3 - 10\,x^2 - 34x + 26 = 0$,

c) $x^4 + 4x^2 - 3x + 3 = 0$.

C. REZOLVAREA NUMERICA A SISTEMELOR DE ECUATII NELINIARE

§1. Metoda lui Newton

Considerăm sistemul

$$(1.1)\qquad \begin{cases} f_1(x_1,x_2,\ldots,x_n) = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1,x_2,\ldots,x_n) = 0 \end{cases}$$

unde $f_i : D\subset R^n \longrightarrow R$ $(i=1,2,\ldots,n)$ sînt n funcţii neliniare (mai exact nu toate liniare). Prin soluţie a sistemului (1.1) înţelegem un element $x^* = {}^t[x_1^*,x_2^*,\ldots,x_n^*] \in R^n$, cu proprietatea că

$$(1.2)\qquad \begin{cases} f_1(x_1^*,x_2^*,\ldots,x_n^*) = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1^*,x_2^*,\ldots,x_n^*) = 0 \end{cases} .$$

Problema existenţei şi unicităţii soluţiei se rezolvă (local) cu ajutorul teoremei de inversare. Intr-adevăr, dacă

$$\frac{D(f_1,f_2,\ldots,f_n)}{D(x_1,x_2,\ldots,x_n)} \neq 0$$

într-o vecinătate V a unui punct $x^0 = {}^t[x_1^0\ x_2^0\ \ldots\ x_n^0] \in R^n$, sistemul

$$f_i(x_1,x_2,\ldots x_n) = y_i \qquad (i=1,2,\ldots,n)$$

se poate rezolva într-o vecinătate $W\subset V$ a lui x^0 şi se obţine

$$x_i=g_i(y_1,y_2,\ldots,y_n) \qquad (i=1,2,\ldots,n).$$

Dacă în particular ${}^t[0,0,\ldots,0]$ aparţine domeniului de existenţă comun funcţiilor $g_i(y_1,y_2,\ldots,y_n)$ $(i=1,2,\ldots,n)$ atunci

$$x_i^*=g_i(0,0,\ldots,0) \qquad (i=1,2,\ldots,n)$$

este unica soluţie a sistemului (1.1) situată în vecinătatea V. Dar teorema de inversare locală nu dă un procedeu efectiv de inversare, ci este o teoremă de existenţă. Din acest motiv, în general nu se poate vorbi de <u>metode directe</u> de rezolvare a sistemului (1.1). In mod obişnuit se folosesc metode iterative de rezolvare (aproximativă) a acestor sisteme. Dintre metodele iterative, metoda cea mai cunoscută este metoda lui Newton, pe care o prezentăm în cele ce urmează.

Deoarece convergenţa unui proces iterativ se justifică uşor cînd sînt îndeplinite condiţiile teoremei de punct fix a lui Banach, mai întîi vom arăta că - în condiţii destul de largi - sistemul (1.1) se poate transforma într-un sistem echivalent (local) cu el, de forma:

(1.3)
$$\begin{cases} x_1 = g_1(x_1,x_2,\ldots,x_n) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = g_n(x_1,x_2,\ldots,x_n) \end{cases}$$

<u>Propoziţia 1.1</u> <u>Fie</u> $f_1,f_2,\ldots,f_n : D\subset R^n \longrightarrow R$, <u>n funcţii de clasă</u> C^2 <u>pe</u> D <u>şi fie</u> $x^* = {}^t[x_1^*\ x_2^*\ \ldots\ x_n^*]$ <u>o soluţie a sistemului</u>

(1.1)
$$\begin{cases} f_1(x_1,x_2,\ldots,x_n) = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1,x_2,\ldots,x_n) = 0 \end{cases}$$

<u>Dacă</u>

(1.4)
$$\frac{D(f_1,f_2,\ldots,f_n)}{D(x_1,x_2,\ldots,x_n)}\bigg|_{x=x^*} \neq 0$$

<u>atunci</u>:

a) <u>există o vecinătate</u> V <u>a lui</u> x^* <u>astfel încît pentru orice</u> $x \in V$ <u>matricea</u> $[\partial f_i/\partial x_j]$ $(i,j=1,2,\ldots,n)$ <u>este inversabilă; vom nota cu</u> $h_{ij}(x_1,x_2,\ldots,x_n)$ $(i,j=1,2,\ldots,n)$ <u>elementele matricei</u> $[\partial f_i/\partial x_j]^{-1}$;

b) <u>funcţiile</u>

(1.5)
$$g_i(x_1,x_2,\ldots,x_n) = x_i - \sum_{j=1}^{n} h_{ij}(x_1,\ldots,x_n) f_j(x_1,\ldots,x_n)$$

$(i=1,2,\ldots,n)$ <u>definite pe</u> V <u>sînt de clasă</u> C^1 <u>pe</u> V <u>şi</u>

(1.6)
$$\frac{\partial g_i}{\partial x_j}(x_1^*,x_2^*,\ldots,x_n^*) = 0 \qquad (i,j=1,2,\ldots,n) ;$$

c) <u>sistemul</u>

(1.7)
$$\begin{cases} x_1 = g_1(x_1,x_2,\ldots,x_n) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = g_n(x_1,x_2,\ldots,x_n) \end{cases}$$

<u>are soluţia</u> x^* <u>şi reciproc, orice soluţie a sistemului</u> (1.7) <u>(situată în</u> V) <u>este şi soluţie a sistemului</u> (1.1) <u>(adică sistemele</u> (1.1) <u>şi</u> (1.7) <u>sînt local echivalente</u>);

d) <u>pentru orice</u> $0<\alpha<1$, <u>fixat, există</u> $r=r(\alpha)>0$, <u>astfel încît</u> $S(x^*,r) = \{x \in R^n ;\ \|x-x^*\|_\sigma \leq r\} \subset D$ <u>şi</u>

(1.8)
$$\max_i \sup_{x\in S(x^*,r)} \Big(\sum_{j=1}^{n} |\partial g_i/\partial x_j|\Big) \leq \alpha < 1 \ .$$

Demonstrație a) Din (1.4) și din continuitatea derivatelor $\frac{\partial f_i}{\partial x_j}$ rezultă că există o vecinătate V a lui x* astfel încît

(1.9) $$\frac{D(f_1,f_2,..,f_n)}{D(x_1,x_2,..,x_n)}(x) \neq 0$$

pentru orice $x \in V \subset R^n$ (V fiind o vecinătate a lui x*). Această condiție ne asigură că matricea $[\partial f_i/\partial x_j]$ poate fi inversată pe V.

b) Elementele $h_{ij}(x_1,x_2,..,x_n)$ sînt cîturi de sume algebrice de produse de derivate de forma $\partial f_i/\partial x_j$ (deci sînt de clasă C^1). In plus

$$\frac{\partial g_i}{\partial x_k} = \delta_{ik} - \sum_{j=1}^{n} \frac{\partial h_{ij}}{\partial x_k} f_j - \sum_{j=1}^{n} h_{ij}\frac{\partial f_j}{\partial x_k}$$

adică

$$\frac{\partial g_i}{\partial x_k}(x^*) = \delta_{ik} - \sum_{j=1}^{n} h_{ij}(x^*)\frac{\partial f_j}{\partial x_k}(x^*) = 0 .$$

c) Egalitățile

$$g_i(x_1^*,x_2^*,..,x_n^*)=x_i^*-\sum_{j=1}^{n}h_{ij}(x_1^*,x_2^*,..,x_n^*)f_j(x_1^*,x_2^*,..,x_n^*)=x_i^*$$

(i=1,2,..,n) ne asigură că x* este soluție a sistemului (1.7). Dacă $\tilde{x} \in V$ este soluție a sistemului (1.7) atunci

$$g_i(\tilde{x}_1,\tilde{x}_2,..,\tilde{x}_n) = \tilde{x}_i$$

și prin urmare $\sum_{j=1}^{n} h_{ij}(\tilde{x})f_j(\tilde{x}) = 0$. Ținînd cont de (1.9) rezultă că $\tilde{x}$ este soluție și pentru sistemul (1.1).

d) Deoarece $\frac{\partial g_i}{\partial x_j}(x^*) = 0$, rezultă că pentru orice $0<\alpha<1$, fixat, există vecinătățile $W_{ij} \subset D$ (i,j=1,2,..,n) ale lui x* astfel încît $\left|\frac{\partial g_i}{\partial x_j}(x)\right|<\frac{\alpha}{n}$, $\forall x \in W_{ij}$. Există atunci $r>0$ ($r=r(\alpha)$) astfel încît $S(x^*,r)\subset W = \bigcap_{i,j=1}^{n} W_{ij} \subset D$ și astfel încît are loc (1.8).*

Propoziția 1.2 Dacă funcțiile $f_i(x_1,x_2,..,x_n)$ (i=1,2,..,n) satisfac ipotezele Propoziției 1.1, aplicația $G:S(x^*,r)\longrightarrow R^n$ definită prin

(1.10) $$G(x_1,x_2,..,x_n)=(g_1(x_1,x_2,..,x_n),...,g_n(x_1,x_2,..,x_n))$$

($g_i(x_1,x_2,..,x_n)$ fiind funcțiile definite prin relațiile (1.5)) este o contracție pe S(x*,r) (r fiind determinat prin propoziția 1.1d).*

Demonstrație Intr-adevăr, deoarece $S(x^*,r)$ este o mulțime convexă există $\xi \in S(x^*,r)$ (aici ξ depinde de x,y și i) astfel încît

$$\|G(x_1,x_2,..x_n)-G(y_1,y_2,..,y_n)\|_m = \max_{1\le i\le n} |g_i(x_1,x_2,...,x_n) - g(y_1,y_2,..,y_n)| = \max_{1\le i\le n}\left|\frac{\partial g_i}{\partial x_1}(\xi)(x_1-y_1)+..+\frac{\partial g_i}{\partial x_n}(\xi)(x_n-y_n)\right| .$$

Prin urmare

$$\|G(x_1,..,x_n)-G(y_1,..,y_n)\|_m \le \max_{1\le i\le n}\sum_{j=1}^{n}\left|\frac{\partial g_i}{\partial x_j}(\xi)\right| |x_j-y_j| \le$$

$$\le \|x-y\|_m \cdot \max_{1\le i\le n}\sum_{j=1}^{n}\left|\frac{\partial g_i}{\partial x_j}(\xi)\right| \le \|x-y\|_m \cdot \max_{1\le i\le n}\ \sup_{x\in S(x^*,r)}\sum_{j=1}^{n}\left|\frac{\partial g_i}{\partial x_j}(x)\right| \le$$

$$\le \alpha\cdot\|x-y\|_m ;$$

deci $G(S(x^*,r))\subset S(x^*,r)$ deoarece $\|G(x)-x^*\|_m = \|G(x)-G(x^*)\|_m \le$ $\le \|x-x^*\|_m \le r$. Cum $0<\alpha<1$, rezultă că G este o contracție pe $S(x^*,r)$.

Teorema 1.3 Fie $f_1,f_2,..,f_n: D\subset R^n \longrightarrow R$, n funcții de clasă C^2 pe D și fie $x^*=(x_1^*,x_2^*,..,x_n^*) \in D$ o soluție a sistemului

(1.1) $$f_i(x_1,x_2,...,x_n) = 0 \qquad (i=1,2,..n)$$

cu proprietatea

$$\frac{D(f_1,f_2,..,f_n)}{D(x_1,x_2,..,x_n)}(x^*) \ne 0 .$$

Considerăm o vecinătate $V\subset D$ a lui x^* astfel încît pentru orice $x\in V$ matricea $\left[\frac{\partial f_i}{\partial x_j}(x)\right]$ să fie inversabilă și notăm cu $h_{ij}(x)$ $(i,j=1,2,..,n)$ elementele inversei acestei matrice (pe V). In aceste condiții există pentru orice $0<\alpha<1$ un $r>0$ $(r=r(\alpha))$ astfel încît $S(x^*,r)\subset D$ și pentru orice $x^{(0)}\in S(x^*,r)$, șirul definit prin

(1.11) $$x^{(0)},\quad x_i^{(k+1)}=x_i^{(k)}-\sum_{j=1}^{n}h_{ij}(x_1^{(k)},..,x_n^{(k)})f_j(x_1^{(k)},..,x_n^{(k)})$$

$(i=1,2,..,n\ ;k=0,1,2,..)$ este convergent la x^*. In plus

$$|x_i^{(k)} - x_i^*| < r\cdot\alpha^k \qquad (i=1,2,..,n)$$

Demonstrație Construim funcțiile $g_i(x_1,x_2,..,x_n)$ $(i=1,2,..,n)$ prin formulele (1.5). Atunci, sistemul

$$x_i = g_i(x_1,x_2,..,x_n) \qquad (i=1,2,..,n)$$

are soluția x^*. In plus, funcția $G:S(x^*,r)\longrightarrow R^n$ definită prin (1.10) este o contracție pe $S(x^*,r)$ (r fiind numărul real pozi-

tiv asociat prin Propoziţia 1.1d unui α fixat, cu $0<\alpha<1$). Aplicînd teorema de punct fix a lui Banach pentru spaţiul metric complet $(S(x^*,r),d)$ (d fiind metrica asociată normei $\| \|_m^-$) cu contracţia G, se stabileşte că şirul

(1,12) $\qquad x^{(0)}\ ,\quad x^{(k+1)} = G(x_1^{(k)},x_2^{(k)},\dots,x_n^{(k)}) \qquad (k=0,1,2,\dots)$

este convergent la x^* şi că $\|x^{(k)}-x^*\|_m^- \leq r\cdot\alpha^k$. Dar şirul (1.12) este chiar şirul (1.11). In plus

$$|x_i^{(k)} - x_i^*| \leq \|x^{(k)} - x^*\|_m^- \leq r\cdot\alpha^k \qquad (i=1,2,\dots,n)\ .*$$

<u>Exemplul 1.1</u> (Demidovici) Să determinăm soluţia cu componente pozitive a sistemului de ecuaţii

$$\begin{cases} f_1(x_1,x_2) = x_1 + 2\lg x_1 - x_2^2 = 0 \\ f_2(x_1,x_2) = 2x_1^2 - x_1x_2 - 5x_1 + 1 = 0\ . \end{cases}$$

Pornind de la aproximaţia iniţială $x^{(0)} = {}^t[3,4 \quad 2,2]$ (determinată folosind reprezentările grafice ale curbelor $f_1(x_1,x_2)=0$ şi $f_2(x_1,x_2)=0$) se găseşte

$$f_1(x_1^0,x_2^0) = 0{,}1544 \qquad f_2(x_1^0,x_2^0) = -\,0{,}3600\ .$$

Atunci matricea

$$\frac{D(f_1,f_2)}{D(x_1,x_2)}(x_1^0,x_2^0) = \begin{bmatrix} 1{,}3832 & -4{,}4 \\ 6{,}4 & -3{,}4 \end{bmatrix}$$

are inversa

$$\frac{1}{23{,}4571}\begin{bmatrix} -3{,}4 & 4{,}4 \\ -6{,}4 & 1{,}3832 \end{bmatrix}\ .$$

Folosind (1.11) se găseşte

$$x^{(1)} = {}^t[3{,}4899 \quad 2{,}2633]$$

Similar se găsesc aproximantele

$$x^{(2)} = {}^t[3{,}4891 \quad 2{,}2621]$$
$$x^{(3)} = {}^t[3{,}4875 \quad 2{,}2616]\ .$$

Pentru $x^{(3)}$ vectorul rezidual este : ${}^t[0{,}002 \quad 0{,}0000]$.*

§2. <u>Metoda gradientului</u>

Considerăm sistemul neliniar

(2.1) $$\begin{cases} f_1(x_1,x_2,\dots,x_n) = 0 \\ \dots\dots\dots\dots\dots\dots \\ f_n(x_1,x_2,\dots,x_n) = 0 \end{cases}$$

pe care-l putem pune sub forma vectorială

(2.2) $$f(x) = 0$$

utilizînd notațiile: $x = {}^t[x_1 \ldots x_n]$ şi $f = {}^t[f_1 \ldots f_n]$.

Presupunem că funcţiile f_i $(i=1,2,\ldots,n)$ sînt de clasă C^1 pe domeniul de definiţie comun $D \subset R^n$. Construim funcţia:

(2.3) $$U(x) = \sum_{i=1}^{n}\left[f_i(x)\right]^2 = (f(x),f(x)) \ .$$

(aici (,) notează produsul scalar în R^n). Rezultă că $x^* \in D$ este soluţie a sistemului (2.2) dacă şi numai dacă $U(x^*)=0$. Presupunem că sistemul (2.1) nu are decît o soluţie izolată în D, care este un punct de minim strict pentru U. In aceste condiţii rezolvarea sistemului (2.1) este echivalentă cu determinarea minimului funcţiei (cîmpului scalar) U(x).

Fie $x^{(o)}$ o aproximaţie a soluţiei x^* a sistemului (2.1). Considerăm suprafaţa de nivel a funcţiei U(x) ce trece prin $x^{(o)}$; ea are ecuaţia

(2.4) $$U(x) = U(x^{(o)}) \ .$$

şi este un elipsoid. Se ştie că grad$U(x)$ este direcţia în care valorile cîmpului scalar U(x) variază cel mai rapid. Ştim totodată că grad$U(x^{(o)})$ dă direcţia normalei în $x^{(o)}$ la suprafaţa (2.4). Pornind din $x^{(o)}$, ne deplasăm pe direcţia normalei la suprafaţa de nivel (2.4) pînă cînd normala devine tangentă într-un punct $x^{(1)}$ la o anumită suprafaţă de nivel. Pornind din $x^{(1)}$ pe normala în punctul $x^{(1)}$ al suprafeţei de cîmp $U(x)=U(x^{(1)})$ ne deplasăm pînă în punctul $x^{(2)}$ în care această normală este tangentă suprafeţei $U(x)=U(x^{(2)})$. Repetînd acest procedeu construim un şir $\{x^{(k)}\}_{k\in N} \subset R^n$ cu proprietatea

(2.5) $$U(x^{(o)}) > U(x^{(1)}) > \ldots > U(x^{(k)}) > \ldots \ .$$

După ce am lămurit aspectele geometrice ale acestei metode să trecem să stabilim aspectele analitice ale metodei. Folosind notaţia

(2.6) $$\operatorname{grad} U(x) = {}^t\left[\frac{\partial U}{\partial x_1}\ \frac{\partial U}{\partial x_2}\ \ldots\ \frac{\partial U}{\partial x_n}\right]$$

rezultă că

(2.7) $$x^{(k+1)} = x^{(k)} - \lambda_k \text{ grad } U(x^{(k)}) \qquad (k=0,1,2,\ldots).$$

Rămîn de determinat factorii λ_k. Considerăm în acest scop funcţia

(2.8) $$F(\lambda) = U(x^{(k)} - \lambda \cdot \text{grad}\, U(x^{(k)}))$$

care dă variaţia funcţiei U în lungul normalei în $x^{(k)}$ la suprafaţa de nivel ce trece prin $x^{(k)}$. Factorul $\lambda = \lambda_k$ trebuie ales astfel încît $F(\lambda)$ să fie minimă, adică trebuie ca $F'(\lambda_k) = 0$. Deci λ_k trebuie să fie rădăcina pozitivă (minimală) a ecuaţiei

(2.9) $$F'(\lambda) = \frac{d}{d\lambda} U(x^{(k)} - \lambda \cdot \text{grad}\, U(x^{(k)})) = 0 .$$

Ecuaţia (2.9) trebuie rezolvată numeric. Deoarece această rezolvare este incomod de realizat pentru fiecare k în parte, în practică se obişnuieşte să se procedeze în felul următor. Se consideră că λ are o valoare suficient de mică aşa fel încît λ^s cu $s \geq 2$ se pot neglija. Dezvoltînd $F(\lambda)$ în serie de puteri în jurul lui $x^{(k)}$ şi oprind numai termenii ce conţin puterile 0 şi 1 ale lui λ, obţinem:

(2.10) $$\tilde{F}(\lambda) = \sum_{i=1}^{n} \left[f_i(x^{(k)}) - \lambda \frac{\partial f_i(x^{(k)})}{\partial x} \text{ grad} U(x^{(k)}) \right]^2$$

unde

$$\frac{\partial f_i}{\partial x} = \left[\frac{\partial f_i}{\partial x_1} \ \frac{\partial f_i}{\partial x_2} \ \cdots \ \frac{\partial f_i}{\partial x_n} \right] .$$

Rezultă

$$\tilde{F}'(\lambda) = -2 \sum_{i=1}^{n} \left[f_i(x^{(k)}) - \lambda \frac{\partial f_i(x^{(k)})}{\partial x} \text{grad} U(x^{(k)}) \right] \frac{\partial f_i}{\partial x}(x^{(k)}) \text{ grad} U(x^{(k)}) .$$

Vom lua ca λ_k soluţia ecuaţiei $\tilde{F}'(\lambda) = 0$, adică

(2.11) $$\lambda_k = \frac{\sum_{i=1}^{n} f_i(x^{(k)}) \left[\frac{\partial f_i}{\partial x}(x^{(k)}) \cdot \text{grad } U(x^{(k)}) \right]}{\sum_{i=1}^{n} \left[\frac{\partial f_i}{\partial x}(x^{(k)}) \text{ grad} U(x^{(k)}) \right]^2}$$

Făcînd notaţia

(2.12) $$W(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

obţinem

(2.13) $$\frac{\partial U}{\partial x_j} = 2 \left\{ {}^t W(x) \cdot f(x) \right\}_j$$

şi

(2.14) $$\text{grad } U(x) = 2 \cdot {}^t W(x) \cdot f(x) .$$

Folosind (2.14) şi (2.11) obţinem

$$(2.15)\qquad \mu_k = 2\lambda_k = \frac{(f^{(k)}, W_k \cdot {}^tW_k f^{(k)})}{(W_k \cdot {}^tW_k f^{(k)}, W_k \cdot {}^tW_k f^{(k)})}$$

unde am folosit notaţiile $W_k = W(x^{(k)})$ şi $f^{(k)} = f(x^{(k)})$.

Relaţia de recurenţă care defineşte şirul aproximaţiilor succesive pentru metoda gradientului este:

$$(2.16)\qquad x^{(k+1)} = x^{(k)} - \mu_k \cdot {}^tW_k f^{(k)} \qquad (k=0,1,2,\ldots).$$

<u>Exemplul 2.1</u>(Demidovici) Să evaluăm, folosind metoda gradientului, soluţia aproximativă a sistemului

$$\begin{cases} x + x^2 - 2yz = 0,1 \\ y - y^2 + 3xz = -0,2 \end{cases}, \quad z + z^2 + 2xy = 0,3$$

situată într-o vecinătate a originii.

Alegem $x^{(0)} = {}^t[0\ 0\ 0]$ şi considerăm matricele:

$$f = \begin{bmatrix} x+x^2-2yz-0,1 \\ y-y^2+3xz+0,2 \\ z+z^2+2xy-0,3 \end{bmatrix} \qquad W(x,y,z) = \begin{bmatrix} 1+2x & -2z & -2y \\ 3z & 1-2y & 3x \\ 2y & 2x & 1+2z \end{bmatrix}.$$

Se găseşte : $f^{(0)} = {}^t[-0,1\quad 0,2\quad -0,3]$ şi $W_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Rezultă că $\mu_0 = 1$ şi $x^{(1)} = x^{(0)} - W_0 f^{(0)} = {}^t[0,1\quad -0,2\quad 0,3]$.

Determinăm

$$f^{(1)} = {}^t[0,13\quad 0,05\quad 0,05],\quad W_1 = \begin{bmatrix} 1,2 & -0,6 & 0,4 \\ 0,9 & 1,4 & 0,3 \\ -0,4 & 0,2 & 1,6 \end{bmatrix}$$

$${}^tW_1 \cdot f^{(1)} = {}^t[0,181\quad 0,002\quad 0,147],$$

$$W_1 \cdot {}^tW_1 \cdot f^{(1)} = {}^t[0,2748\quad 0,2098\quad 0,1632] .$$

Evaluăm μ_1 cu ajutorul formulei (2,15):

$$\mu_1 = \frac{0,13\cdot 0,2748 + 0,05\cdot 0,2098 + 0,05\cdot 0,1632}{0,2748^2 + 0,2098^2 + 0,1632^2} = \frac{0,054374}{0,14619797} = 0,3719.$$

Obţinem aproximanta $x^{(2)}$ şi rezidualul $f^{(2)}$ prin

$$x^{(2)} = \begin{bmatrix} 0,1 \\ -0,2 \\ 0,3 \end{bmatrix} - 0,3719 \begin{bmatrix} 0,181 \\ 0,002 \\ 0,147 \end{bmatrix} = \begin{bmatrix} 0,0327 \\ -0,2007 \\ 0,2453 \end{bmatrix},\quad f^{(2)} = \begin{bmatrix} 0,032 \\ -0,017 \\ -0,007 \end{bmatrix}.$$

<u>Observaţia 2.1</u> Metoda gradientului se poate aplica şi sistemelor liniare $AX = b$, luînd $f = AX - b$. In acest caz se obţine $W = A$.

§3. Exerciţii

1. Folosind metoda lui Newton şi metoda gradientului să se determine soluţia aproximativă cu componente pozitive a sistemului

$$\begin{cases} x^2 + y^2 + z^2 = 1 \\ 2x^2 + y^2 - 4z = 0 \\ 3x^2 - 4y + z^2 = 0 \end{cases}.$$

Utilizaţi $x^{(0)} = {}^t[0,5 \quad 0,5 \quad 0,5]$.

2. Folosind metoda lui Newton şi metoda gradientului să se determine soluţia aproximativă a sistemului

$$\begin{cases} x^2 + y^2 = 1 \\ x^3 - y = 0 \end{cases}$$

a) cu ambele componente pozitive,

b) cu ambele componente negative.

(Pentru a alege $x^{(0)}$ utilizaţi graficele celor două curbe.)

3. Folosind metoda gradientului să se determine soluţia aproximativă a sistemului

$$\begin{cases} 11x_1 + 4x_2 + 3x_3 - x_4 = -1 \\ 2x_1 - 12x_2 + x_3 + 3x_4 = -22 \\ 3x_1 - x_2 + 9x_3 + 2x_4 = 6 \\ x_1 + 2x_2 - 3x_3 + 8x_4 = 8 \end{cases}.$$

CAPITOLUL IV

METODE NUMERICE DE DETERMINARE A VALORILOR PROPRII

§1. Generalităţi

Calculul valorilor proprii şi al vectorilor proprii pentru o matrice pătratică dată este una dintre cele mai dificile probleme ale algebrei liniare. Problema valorilor proprii se rezolvă de obicei prin metode iterative, datorită faptului că valorile proprii sînt de fapt rădăcinile unui polinom algebric. Dar determinarea polinomului caracteristic cu ajutorul calculatorului electronic este afectată de erori. Există exemple care arată că modificări foarte mici ale coeficienţilor polinomului caracteristic conduc la polinoame ale căror rădăcini sînt mult diferite de rădăcinile polinomului iniţial. Mai exact, există exemple în care, dacă un singur coeficient este apreciat cu o eroare mai mică chiar decît eroarea de rotunjire acceptată, rădăcinile polinomului aproximant sînt mult diferite de rădăcinile polinomului caracteristic în discuţie. Unul dintre cele mai cunoscute exemple de acest fel este exemplul lui J.H.Wilkinson, pe care îl prezentăm în cele ce urmează.

Exemplul lui Wilkinson. Se consideră matricea de ordinul 20

$$A = \begin{bmatrix} 20 & 2 & 0 & \cdots & 0 & 0 \\ 0 & 19 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 2 & 2 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

Polinomul ei caracteristic este $(\lambda-1)(\lambda-2)\ldots(\lambda-20)$. Dacă pe locul $(20,1)$ punem ε în locul lui 0, polinomul caracteristic al matricei obţinute este $(\lambda-1)(\lambda-2)\ldots(\lambda-20) = 2^{19}\varepsilon$. Dacă $\varepsilon = 10^{-10}$, 14 valori proprii sînt complexe, cu părţile imaginare mai mari ca 1.*

Acesta este motivul pentru care se preferă să se determine valorile proprii ca limite ale unor şiruri numerice,fără determinarea prealabilă a coeficienţilor polinomului caracteristic. De obicei se caută simultan şi valorile proprii şi vectorii proprii corespunzători (sau alţi vectori legaţi de vectorii proprii prin relaţii simple).Cele mai bune rezultate se obţin pentru matricele simetrice (sau pentru cele hermitiene).Determinarea valorilor proprii şi a vectorilor proprii pentru astfel de matrice este echivalentă cu determinarea axelor unor elipsoizi.In cazul matricelor nesimetrice,problema se rezolvă mult mai greu.

Fie $\mathcal{M}_n(K)$ - mulţimea matricelor pătratice,de ordinul n,cu coeficienţi din cîmpul K.In cele ce urmează K va fi R sau C.Ca şi pînă acum,vom considera K^n identificat natural cu mulţimea $\mathcal{M}_{nx1}(K)$ a matricelor coloană,cu n elemente din K.

<u>Definiţia 1.1</u> <u>Vectorul nenul</u> $u \in K^n$ <u>se numeşte vector propriu pentru matricea</u> $A \in \mathcal{M}_n(K)$ <u>dacă există un scalar</u> $\lambda \in K$,<u>numit valoare proprie pentru</u> A,<u>astfel încît</u>

(1.1) $$A u = \lambda \cdot u \quad .$$

<u>Observaţia 1.1</u> 1°. Definiţia 1.1 nu exclude posibilitatea $\lambda = 0 \in K$.Vectorii proprii corespunzători valorii proprii $\lambda = 0$ formează un subspaţiu vectorial al lui K^n numit <u>spaţiul nul</u> al matricei A.

2°. Dacă $u \in K^n$ este vector propriu al matricei A,corespunzător valorii proprii λ ,are loc şi egalitatea

(1.1') $$^t u \cdot {}^t A = \lambda \cdot {}^t u \quad .$$

Prin urmare se poate pune problema definirii valorilor şi vectorilor proprii la dreapta (prin definiţia 1.1) şi a vectorilor şi valorilor proprii la stînga,prin

(1.2) $$^t v \cdot A = \lambda \cdot {}^t v \quad .$$

Dacă $A = {}^t A$,cele două definiţii conduc la aceleaşi valori proprii şi aceiaşi vectori proprii.

3°. Problema determinării valorilor şi vectorilor proprii ai unei matrice A se poate reformula şi astfel: să se determine valorile $\lambda \in K$ astfel încît sistemul liniar şi omogen (1.1) să aibă soluţii netriviale. Folosind teoria sistemelor liniare va rezulta că valorile $\lambda \in K$, soluţii ale acestei probleme, sînt rădăcinile polinomului

$$\det(A - \lambda I_n) = 0, \tag{1.3}$$

numit <u>polinomul caracteristic</u> al matricei A.

Deoarece ${}^t(A - \lambda I_n) = {}^tA - \lambda I_n$, rezultă că

$$\det(A - \lambda I_n) = \det({}^tA - \lambda I_n)$$

şi deci, valorile proprii ale matricei A coincid cu cele ale matricei ${}^tA$. Vectorii proprii asociaţi matricelor A şi ${}^tA$ vor fi în general diferiţi. Intr-adevăr, să presupunem că matricea A are valorile proprii $\lambda_1, \lambda_2, \ldots, \lambda_n$ (nu neaparat distincte) şi că există o bază formată din vectorii proprii $\{e_1, e_2, \ldots, e_n\}$ corespunzători acestor valori proprii. Fie B - matricea care are pe coloane componentele vectorilor $e_1, e_2, \ldots$, şi respectiv e_n. Vom scrie

$$B = [e_1 \ e_2 \ \ldots \ e_n].$$

Egalităţile $Ae_i = \lambda_i e_i$ se pot prinde în egalitatea matriceală

$$A \cdot B = B \cdot \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

care este echivalentă cu

$$B^{-1}AB = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n).$$

Din ultima egalitate obţinem

$$B^{-1}A = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \cdot B^{-1},$$

deci liniile i $(1 \leq i \leq n)$ ale matricei B^{-1} vor fi vectori proprii la stînga pentru A, corespunzători valorilor proprii λ_i. Dacă $B^{-1} = {}^t[f_1 \ f_2 \ \ldots \ f_n]$, atunci $f_i A = \lambda_i f_i$ şi ${}^tA \cdot {}^tf_i = \lambda_i {}^tf_i$. Deci coloanele matricei B sînt vectori proprii pentru A, iar liniile matricei B^{-1} sînt vectori proprii pentru ${}^tA$.

4°. Egalităţile

$$\det(S^{-1}AS - \lambda I_n) = \det(S^{-1}AS - \lambda S^{-1}I_n S) = \det(A - \lambda I_n)$$

asigură că matricele A şi $S^{-1}AS$ au aceleaşi valori proprii.

§2. Localizarea valorilor proprii

Fie $A \in \mathcal{M}_n(K)$ cu $A = [a_{ij}]$. Polinomul caracteristic al matricei A se poate scrie sub forma

$$(2.1)\qquad \det(A-\lambda I_n) = (-1)^n[\lambda^n - \alpha_{n-1}\lambda^{n-1} + \ldots + (-1)^{n-1}\alpha_1\lambda + (-1)^n\alpha_n] \equiv p(\lambda)$$

unde

$$\alpha_{n-k} = \sum_{\substack{i_1,i_2,\ldots,i_k=1 \\ i_1<i_2<\ldots<i_k}}^{n} \det\begin{bmatrix} i_1 & i_2 \ldots i_k \\ i_1 & i_2 \ldots i_k \end{bmatrix} \qquad (k=1,2,\ldots,n)$$

Dacă $\lambda_1, \lambda_2, \ldots, \lambda_n \in C$ sînt valorile proprii ale matricei A, se stabilește imediat (folosind ultima relație Viète pentru polinomul (2.1)) că cel puțin una dintre rădăcinile polinomului caracteristic se află în discul $|\lambda| \le |\det A|^{1/n}$. Folosind inegalitatea lui Hadamard ($|\det A| \le \prod_{i=1}^{n} |C_i|$, unde C_i notează lungimea vectorului coloană C_i) pentru determinanții matricelor A și ${}^tA$ se stabilește că cel puțin o valoare proprie a matricei A se găsește în fiecare dintre dicurile

$$(2.2)\qquad |\lambda| \le \max_{1\le i\le n}\Big(\sum_{j=1}^{n}|a_{ij}|^2\Big)^{\frac12}, \quad |\lambda| \le \max_{1\le j\le n}\Big(\sum_{i=1}^{n}|a_{ij}|^2\Big)^{\frac12}.$$

Alte estimări ale domeniilor ce conțin valori proprii se pot obține cu ajutorul normelor matriceale compatibile cu norme vectoriale. Intr-adevăr, pentru orice normă matriceală $\|\ \|$ compatibilă cu norma vectorială $\|\ \|_v$, are loc inegalitatea

$$\|Ax\|_v \le \|A\|\cdot\|x\|_v .$$

(In continuare vom nota cu același simbol și norma vectorială și norma matriceală indusă.)

Dacă u este vector propriu corespunzător valorii proprii λ, are loc relația $Au = \lambda u$ și prin urmare are loc inegalitatea

$$|\lambda|\cdot\|u\| = \|Au\| \le \|A\|\cdot\|u\|.$$

Cum $u \ne 0$, rezultă $|\lambda| \le \|A\|$. Similar găsim $|\lambda| \le \|{}^tA\|$. Pentru $\|\ \|_1$ și $\|\ \|_\infty$ aceste inegalități devin

$$(2.3)\qquad |\lambda| \le \max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}| \qquad |\lambda| \le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}| .$$

Estimarea obținută pe această cale cu ajutorul normei lui Frobe-

nius (euclideană) este slabă deoarece are loc inegalitatea

$$\sum_{i=1}^{n}|\lambda_i|^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2 \quad .$$

Am obținut în acest fel cîteva procedeie de localizare a valorilor proprii. Dar, deși sînt suficient de bune în unele aplicații, aceste procedeie au următoarele dezavantaje:

a) estimarea este globală (pentru toate valorile proprii)

b) estimarea nu este suficient de fină.

Aprecieri ceva mai comode și mai bune se obțin pe baza discurilor Gerschgorin.

Dovedim mai întîi

<u>Lema 2.1</u> <u>Pentru ca matricea A să fie nesingulară este suficient să fie satisfăcute inegalitățile:</u>

(2.4)
$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| \qquad (i=1,2,\ldots,n).$$

<u>Demonstrație</u> Presupunem că, deși au loc relațiile (2.4), matricea A are $\det A = 0$. Atunci sistemul de ecuații liniare

$$\sum_{j=1}^{n} a_{ij}x_j = 0 \qquad (i=1,2,\ldots,n)$$

are o soluție nebanală $(x_1^0, x_2^0, \ldots, x_n^0)$. Fie x_k^0 componenta de modul maxim a acestei soluții. Atunci ecuația de rang k a sistemului conduce la egalitatea

$$a_{kk}x_k^0 = -\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j^0 \quad .$$

De aici se obțin inegalitățile

$$|a_{kk}|\cdot|x_k^0| \leq \sum_{\substack{j=1\\ j\neq k}}^{n}|a_{kj}|\cdot|x_j^0| \leq \sum_{\substack{j=1\\ j\neq k}}^{n}|a_{kj}|\cdot|x_k^0| \ ,$$

adică

$$|a_{kk}| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj} \quad .$$

Această relație contrazice ipoteza (2.4), deci presupunerea că $\det A = 0$ este falsă.*

<u>Definiția 2.1</u> <u>Dacă matricea</u> $A = [a_{ij}] \in \mathcal{M}_n(K)$ <u>satisface condiția (2.4) se spune că are diagonala dominantă pe linii; dacă matricea A satisface inegalitățile</u>

(2.4')
$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ji}| \qquad (i=1,2,\ldots,n)$$

<u>se spune că are diagonala dominantă pe coloană.</u>

Corolar 2.1 Pentru ca λ să fie valoare proprie a matricei A este necesar să existe cel puţin un indice i $(1\le i\le n)$ astfel încît să fie satisfăcută inegalitatea

(2.5) $$|\lambda - a_{ii}| \le \sum_{\substack{j=1\\ j\ne i}}^{n} |a_{ij}|$$

adică, orice valoare proprie a matricei A se găseşte cel puţin în unul dintre discurile de centre a_{ii} şi raze $\sum_{\substack{j=1\\ j\ne i}}^{n} |a_{ij}|$ $(1\le i\le n)$.*

Demonstraţie Afirmaţia rezultă direct din lema 2.1 observînd că pentru orice valoare proprie λ a matricei A, matricea B = = $A - \lambda I_n$ are determinantul egal cu zero.

Definiţia 2.2 Domeniile definite prin relaţiile (2.5) se numesc discuri Gerschgorin ale matricei A.*

Deoarece A şi ${}^tA$ au aceleaşi valori proprii, aplicînd Corolarul 2.1 matricei ${}^tA$ rezultă că orice valoare proprie a matricei A se găseşte cel puţin în unul dintre discurile

(2.6) $$|\lambda - a_{ii}| \le \sum_{\substack{j=1\\ j\ne i}}^{n} |a_{ji}| \ .$$

Definiţia 2.2') Domeniile definite prin relaţiile (2.6) se numesc tot discuri Gerschgorin ale matricei A.*

Folosind discurile Gerschgorin se pot izola anumite valori proprii. Are loc

Teorema 2.1 Dacă s discuri Gerschgorin formează un domeniu (conex) G izolat de celelalte discuri, atunci domeniul G conţine exact s valori proprii ale matricei A.*

Demonstraţie Demonstraţia se bazează pe dependenţa continuă a valorilor proprii ale unei matrice de elementele matricei. Punem matricea A sub forma A = B + C, unde $B = \mathrm{diag}(a_{11},\dots,a_{nn})$. Considerăm familia de matrice $A_\varepsilon = B + \varepsilon C$ unde $0\le\varepsilon\le 1$. Comparînd discurile Gerschgorin de acelaşi centru a_{ii} ale matricelor A şi A_ε se remarcă că razele lor se disting prin factorul multiplicativ ε. Să notăm cu G_ε domeniul închis obţinut reunind discurile Gerschgorin ale matricei A_ε de centre aparţinînd lui G. Prin F_ε notăm domeniul închis format din reuniunea celorlalte

discuri. Au loc relațiile $G_\varepsilon \subset G$ și $G_\varepsilon \cap F_\varepsilon = \emptyset$ pentru $0 \le \varepsilon \le 1$. Pentru $\varepsilon = 0$, domeniul G_0 conține exact s valori proprii ale matricei A_0. Valorile proprii variază continuu ca funcții de ε. Să presupunem că pentru un anumit ε una dintre valorile proprii iese din domeniul G_ε. Atunci, în virtutea continuității și a condiției $G_\varepsilon \cap F_\varepsilon = \emptyset$ ar rezulta că există un ε astfel încît una dintre valorile proprii ale matricei A_ε nu aparține nici lui G_ε, nici lui F_ε. Acest lucru este imposibil; deci pentru toate valorile admisibile ale lui ε, G_ε conține exact s valori proprii (distincte sau nu) ale matricei A_ε. Dar pentru $\varepsilon=1$, G_ε coincide cu G, iar matricea A_ε coincide cu A, deci G conține s valori proprii.

Corolarul 2.2 Dacă un disc Gerschgorin oarecare este izolat, el conține exact o valoare proprie.*

Corolarul 2.3 Dacă pentru un anumit i (fixat) au loc inegalitățile

(2.7) $$|a_{kk} - a_{ii}| > \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}| + \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{ij}|$$

(k=1,2,..,i-1,i+1,..,n) atunci discul Gerschgorin

$$|\lambda - a_{ii}| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$$

conține exact o valoare proprie.*

Să considerăm matricea

(2.8) $$B = DAD^{-1}$$

unde $D = \mathrm{diag}(d_1, d_2, .., d_n)$. Atunci $B = \left[a_{ij}\frac{d_i}{d_j}\right]$. Cum A și B au aceleași valori proprii, rezultă că orice valoare proprie a lui A se găsește în unul dintre discurile

(2.9) $$|\lambda - a_{ii}| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \left|\frac{d_i}{d_j}\right| .$$

Alegînd convenabil matricea D se pot schimba razele discurilor Gerschgorin, se pot izola discuri sau grupe de discuri, adică se poate face o localizare mai exactă a valorilor proprii ale matricei A.

§3. Metoda lui Danilevski de reducere a unei matrice la forma Frobenius

Definiţia 3.1 Se spune că matricea $A \in \mathcal{M}_n(K)$ a fost redusă la forma Frobenius B, unde

$$(3.1)\qquad B = \begin{bmatrix} 0 & 1 & 0 & 0 \dots 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \dots 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 \dots 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \dots 0 & 0 & 1 \\ b_{n1} & b_{n2} & b_{n3} & b_{n4} \dots b_{n,n-2} & b_{n,n-1} & b_{n,n} \end{bmatrix}$$

dacă s-a determinat o matrice nesingulară $S \in \mathcal{M}_n(K)$ astfel încît $S^{-1}AS = B$.*

Propoziţia 3.1 Polinomul caracteristic al matricei B este:

$$(3.2)\qquad p(\lambda)=(-1)^n(\lambda^n-b_{nn}\lambda^{n-1}-b_{n,n-1}\lambda^{n-2}-\dots-b_{n2}\lambda - b_{n1})$$

Demonstraţie Dezvoltînd $\det(B-\lambda I_n)$ după ultima linie obţinem:

$$\det(B-\lambda I_n)=(b_{nn}-\lambda)(-\lambda)^{n-1}-b_{n,n-1}(-\lambda)^{n-2}+b_{n,n-2}(-\lambda)^{n-3}-\dots+(-1)^{n+1}b_{n1}=$$

$$=(-1)^n(\lambda^n-b_{nn}\lambda^{n-1}-b_{n,n-1}\lambda^{n-2}-b_{n,n-2}\lambda^{n-3}-\dots-b_{n2}\lambda - b_{n1}) \text{ .*}$$

Vom arăta în cele ce urmează că orice matrice pătratică A se poate aduce la forma Frobenius cu ajutorul unei matrice S care este produs de transformări elementare ale matricei A.

La primul pas construim matricele $M^{(1)} = m^{(1)}_{ij}$, unde $m^{(1)}_{ij}=\delta_{ij}$ pentru $i\neq 2$ şi $m^{(1)}_{2j}=-a_{1j}/a_{12}$ dacă $a_{12}\neq 0$. Inversa matricei $M^{(1)}$ este matricea $P^{(1)}= p^{(1)}_{ij}$ unde $p^{(1)}_{ij}=\delta_{ij}$ pentru $i\neq 2$ şi $p^{(1)}_{2j}=a_{1j}$. Matricea $A^{(2)}=P^{(1)}A^{(1)}M^{(1)}= \left[a^{(2)}_{ij}\right]$ are $a^{(2)}_{12}=1$ şi $a^{(2)}_{1j}=0$ pentru $j\neq 2$ (aici $A^{(1)}=A$). Intr-adevăr, au loc egalităţile (pentru $j\neq 2$)

$$a^{(2)}_{1j}=\sum_{k=1}^{n}(\sum_{s=1}^{n}p^{(1)}_{1s}a_{sk})m_{kj}=(\sum_{s=1}^{n}p^{(1)}_{1s}a_{s2})m_{2j}+\sum_{s=1}^{n}p^{(1)}_{1s}a_{sj}=a_{12}m_{2j}+a_{1j}=0;$$

pentru $j=2$ au loc egalităţile

$$a^{(2)}_{12}=\sum_{k=1}^{n}(\sum_{s=1}^{n}p^{(1)}_{1s}a_{sk})m_{k2}=\sum_{k=1}^{n}a_{1k}m^{(1)}_{k2}=a_{12}m^{(1)}_{22}=a_{12}\cdot\frac{1}{a_{12}} = 1.$$

Dacă $a^{(2)}_{23}\neq 0$ construim matricele $M^{(2)}=\left[m^{(2)}_{ij}\right]$ şi $P^{(2)}=\left[p^{(2)}_{ij}\right]$ inversa una alteia prin: $m^{(2)}_{ij}=\delta_{ij}$ pentru $i\neq 3$ şi $m^{(2)}_{3j}=-a^{(2)}_{2j}/a^{(2)}_{23}$ şi respectiv $p^{(2)}_{ij}=\delta_{ij}$ pentru $i\neq 3$ şi $p^{(2)}_{3j}=a^{(2)}_{2j}$. Cu ajutorul lui $M^{(2)}$ şi $P^{(2)}$ se construieşte $A^{(3)} = P^{(2)}A^{(2)}M^{(2)}$.

Procesul se repetă de n-1 ori dacă $a^{(i)}_{i,i+1} \neq 0$ $(i=1,2,..,n-1)$. Dacă unul dintre elementele $a^{(i)}_{i,i+1}$ este zero, schimbînd coloana i+1 cu o coloană $j > i+1$ și apoi scimbînd între ele liniile i+1 și j (pentru a obține o matrice asemenea cu matricea precedentă) se poate obține o matrice $A'^{(i)}$ cu elementul $a'^{(i)}_{i,i+1} \neq 0$. In cazul în care nici astfel de schimbări nu ajută, matricea A se transformă într-o matrice diagonală cu blocuri, blocurile fiind de forma Frobenius.

Determinarea unui vector propriu $u = {}^t[u_1\ u_2 \ldots u_n]$ corespunzător valorii proprii λ (pentru matricea B) se face rezolvînd sistemul

$$\begin{cases} -\lambda u_1 + u_2 & = 0 \\ \quad -\lambda u_2 + u_3 & = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ \quad\quad -\lambda u_{n-1} + u_n & = 0 \\ b_{n1} u_1 + b_{n2} u_2 + \ldots + b_{n,n-1} u_{n-1} + (b_{nn} - \lambda) u_n & = 0 \end{cases}$$

Luînd $u_1 = 1$, rezultă $u_2 = \lambda$; $u_3 = \lambda^2, \ldots, u_n = \lambda^{n-1}$. Ultima ecuație joacă rol de ecuație de control. Dacă toate calculele sînt exacte înlocuind u_i cu λ^{i-1} $(i=1,2,..,n)$ în ultima ecuație, obținem egalitate în virtutea faptului că λ este valoare proprie. Dacă u este determinat, atunci Su este vectorul propriu al matricei A corespunzător valorii proprii λ.

§4. Metoda lui Jacobi de determinare a valorilor proprii pentru matrice simetrice

Se știe că pentru orice matrice pătratică simetrică, reală $A = {}^tA \in \mathcal{M}_n(R)$ există o matrice ortogonală S astfel încît

$$(4.1) \qquad {}^tSAS = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

unde $\lambda_i \in R$ $(i=1,2,..,n)$ sînt valorile proprii (distincte sau nu) ale matricei A, iar coloanele lui S sînt coordonatele vectorilor proprii în baza canonică.

Prin metoda lui Jacobi matricea A este transformată într-o matrice diagonală cu ajutorul unui șir de matrice de rotație $\{R_k\}$

Să notăm $A_0 = A$. Definim inductiv şirul $\{A_k\}_{k\in N}$ prin

(4.2) $$A_k = R_k A_{k-1}\,{}^tR_k \qquad (k=1,2,\dots,).$$

Inductiv se arată că matricele A_k sînt simetrice pentru $k=0,1,2,3,\dots$ .Fie $a_{pq}^{(k-1)}$ elementul de modul maxim al matricei A_{k-1} situat deasupra diagonalei principale. Alegem rotaţia R_k în planul (p,q) astfel încît elementul $a_{pq}^{(k)}$ să fie nul. Cum matricea rotaţiei R_k se obţine înlocuind minorul $\begin{pmatrix} p & q \\ p & q \end{pmatrix} I_n$ al matricei I_n cu matricea $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, adică

(4.3) $$R_k = \begin{bmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & \cos\theta & \dots & -\sin\theta & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & \sin\theta & \dots & \cos\theta & \dots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow p \\ \\ \leftarrow q \\ \\ \\ \end{matrix};$$

(coloanele marcate p şi q)

rezultă că matricea A_k diferă de A_{k-1} numai prin liniile şi coloanele de rang p şi q. Prin calcul direct se stabileşte că

(4.4) $$a_{pq}^{(k)} = (a_{qq}^{(k-1)} - a_{pp}^{(k-1)})\cdot\cos\theta\sin\theta + a_{pq}^{(k-1)}(\cos^2\theta - \sin^2\theta) = a_{qp}^{(k)}.$$

Impunînd condiţia $a_{pq}^{(k)} = 0$ rezultă, în ipoteza $a_{pp}^{(k-1)} \neq a_{qq}^{(k-1)}$

(4.5) $$\operatorname{tg} 2\theta = 2a_{pq}^{(k-1)}/(a_{pp}^{(k-1)} - a_{qq}^{(k-1)}).$$

Se alege soluţia ecuaţiei (4.5) care satisface condiţia $|\theta| \leq \frac{\pi}{4}$. Dacă $a_{pp}^{(k-1)} = a_{qq}^{(k-1)}$ luăm $\theta = \pm\frac{\pi}{4}$, după cum semnul lui $a_{pq}^{(k-1)}$ este pozitiv sau negativ. De obicei nu avem nevoie de θ ci numai de $\cos\theta$ şi $\sin\theta$. Făcînd notaţiile

(4.6) $$C = 2a_{pq}^{(k-1)} \qquad D = a_{pp}^{(k-1)} - a_{qq}^{(k-1)} \qquad F = C/\sqrt{C^2 + D^2}$$

se găseşte

(4.7) $$\sin\theta = F/\sqrt{2(1+\sqrt{1-F^2})} \qquad \cos\theta = \sqrt{1 - \sin^2\theta}.$$

Are loc

Teorema 4.1 Şirul matricelor $\{A_k\}_{k\in N}$ construit cu metoda lui Jacobi este convergent şi $\lim\limits_{k\to\infty} A_k = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ cu λ_i $(1 \leq i \leq n)$ - valorile proprii ale matricei A.*

Nu prezentăm demonstraţia acestui rezultat.

Dacă ultima rotaţie care se efectuează este R_s, atunci, dacă am lucra fără erori, ar trebui ca

$$(R_s \dots R_2R_1)A({}^tR_1\,{}^tR_2 \dots {}^tR_s) = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$$

adică produsul ${}^t(R_s \dots R_1)$ este o matrice ale cărei coloane sînt vectorii proprii ai matricei A. Dar în practică, elementele sînt calculate aproximativ, deci se poate considera că vectorii coloană ai transpusei ${}^t(R_s \dots R_2R_1)$ sînt vectorii proprii ai matricei A.

Există cîteva dezavantaje ale acestei metode. Unul dintre ele constă în faptul că dacă la o anumită etapă a fost "anulat" un element nesituat pe diagonala principală, la o rotaţie viitoare acest element poate să se transforme într-un element diferit de zero. Apare întrebarea : Cum se poate face că odată transformat în zero un element nesituat pe diagonala principală, toate transformările viitoare nu-l mai modifică?

J.W.Givens a observat că putem reduce orice matrice simetrică la o matrice triplu-diagonală simetrică (iar o matrice oarecare la aşa-numita formă Hessenberg) anulînd elementele în ordinea:

$$\begin{array}{cccc} (1,3) & (1,4) & \dots & (1,n) \\ & (2,4) & \dots & (2,n) \\ & \dots & \dots & \dots \\ & & & (n-2,n) \end{array}$$

cu ajutorul rotaţiilor care implică liniile şi coloanele corespunzătoare. Deci, reducerea la forma tridiagonală a unei matrice simetrice (reale) se realizează cu $\frac{1}{2}(n-1)(n-2)$ rotaţii bidimensionale.

Am văzut în Cap.III, §2 că este uşor de calculat determinantul unei matrice tridiagonale. Este deci uşor de calculat $p(x)$ pentru un x fixat. De aici rezultă următoarea procedură de aproximare a valorilor proprii:

- se obţin margini ale valorilor proprii (de exemplu cu ajutorul discurilor Gerschgorin)
- bisectînd succesiv intervalul în care se află valoarea proprie se obţin aproximări ale valorilor proprii.

Mai exact, dacă valoarea proprie $\lambda_n \in (a,b)$, adică $p(a)p(b) < 0$, se calculează $p(\frac{a+b}{2})$; dacă $p(a)p(\frac{a+b}{2}) < 0$ rădăcina se află în intervalul $(a,\frac{a+b}{2})$, iar dacă $p(\frac{a+b}{2})p(b) < 0$, rădăcina se află în intervalul $(\frac{a+b}{2}, b)$. Noul interval în care se află rădăcina se bisectează din nou, ş.a.m.d..

Numărul de operaţii necesar reducerii la forma tridiagonală prin metoda lui Givens este de circa $\frac{4}{3}n^3$.

O modificare semnificativă a acestui proces de reducere a matricei la forma tridiagonală este dată de Householder. In timp ce procedeul lui Givens realizează reducerea anulînd cîte un singur element la fiecare pas, procedeul lui Householder anulează n-1 elemente ale liniei la fiecare transformare. Numărul de operaţii necesare în utilizarea metodei lui Householder este aproape jumătate din cel necesar metodei lui Givens.

Pentru a prezenta procedeul lui Householder este util să dovedim mai întîi

Lema 4.1 Dacă u este versor, adică ${}^t u\cdot u = 1$, atunci $I_n - 2u\cdot{}^t u$ este matrice ortogonală.

Demonstraţie Au loc egalităţile

$${}^t(I_n-2u\cdot{}^t u)(I_n-2u\cdot{}^t u)=(I_n-2u\cdot{}^t u)^2=I_n-4u\cdot{}^t u+4(u\cdot{}^t u)^2=I_n$$

deoarece $(u\cdot{}^t u)(u\cdot{}^t u)=u({}^t u\cdot u){}^t u=u\cdot{}^t u$.*

Fie $u_2={}^t[0\ b_2\ b_3\ \dots\ b_n]$, astfel încît ${}^t u_2\cdot u_2=1$. Atunci:

$$P_2 = I_n - 2\begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & b_2^2 & b_2b_3 & \dots & b_2b_n \\ 0 & b_3b_2 & b_3^2 & \dots & b_3b_n \\ \dots & \dots & \dots & \dots & \dots \\ 0 & b_nb_2 & b_nb_3 & \dots & b_n^2 \end{bmatrix} \tag{4.8}$$

şi $AP_2=A-2Au_2\cdot{}^t u_2$. Prima linie a lui AP_2 este identică cu prima linie a matricei P_2AP_2 şi este:

$$\left[a_{11}\quad a_{12}-2p_1b_2\quad a_{13}-2p_1b_3\ \dots\ a_{1n}-2p_1b_n\right]$$

unde $p_1=\sum_{i=2}^{n} a_{1i}b_i$ (şi în general $p_j=\sum_{i=2}^{n} a_{ji}b_i$). Dorim ca

$$a_{1j} - 2p_1b_j = 0 \qquad (j=3,4,\dots,n) . \tag{4.9}$$

Ţinînd cont că

$$\sum_{j=1}^{n} a_{1j}^2 = a_{11}^2 + (a_{12} - 2p_1 b_2)^2 \tag{4.10}$$

(deoarece printr-o transformare ortogonală lungimea vectorului avînd drept componente elementele primei linii rămîne neschimbată) şi notînd cu S

$$S = \sum_{j=2}^{n} a_{1j}^2 \tag{4.11}$$

rezultă

$$a_{12} - 2p_1 b_2 = \pm\sqrt{S} \quad . \tag{4.12}$$

Inmulţind relaţiile (4.9) cu b_j $(j=3,4,\ldots,n)$, (4.12) cu b_2 şi sumînd rezultatele, găsim

$$p_1 - 2p_1(b_2^2+\ldots+b_n^2) = \pm\, b_2\sqrt{S}$$

adică

$$p_1 = \mp\, b_2\sqrt{S} \; . \tag{4.13}$$

Semnul lui p_1 se alege prin condiţia $p_1 a_{12} > 0$, în scopul de a obţine pentru b_2 o valoare cît mai mare. Cu p_1 determinat prin relaţia (4.13) revenim la (4.12) şi (4.9) pentru a determina componentele $b_2, b_3, \ldots, b_n$ ale versorului u_2:

$$b_2^2 = \tfrac{1}{2}\left(1 \mp \frac{a_{12}}{\sqrt{S}}\right) \qquad b_j = \frac{\mp a_{1j}}{2b_2\sqrt{S}} \quad (j=3,4,\ldots,n). \tag{4.14}$$

Acesta este primul pas în reducerea matricei la forma tridiagonală (resp. la forma Hessenberg) prin metoda lui Householder.

Pentru pasul k se foloseşte matricea $P_k = I_n - 2u_k \cdot {}^t u_k$, unde $u_k = {}^t[0 \,..\, 0 \; x_k \; x_{k+1} \,..\, x_n]$ cu ${}^t u_k \cdot u_k = 1$, construindu-se matricea

$$P_k P_{k-1} \ldots P_2 A P_2 \ldots P_{k-1} P_k \; .$$

La pasul k=n-1 se obţine forma tridiagonală pentru matricea simetrică (resp. forma Hessemberg pentru o matrice oarecare). Folosind apoi metoda bisecţiei, aşa cum s-a precizat în cazul metodei lui Givens, se pot aproxima valorile proprii ale matricei A. Metoda bisecţiei dă rezultate bune în cazul în care s-a făcut în prealabil separarea rădăcinilor. Un procedeu de separare a rădăcinilor se bazează pe utilizarea unui şir Sturm asociat polinomului caracteristic al unei matrice tridiagonale.

Fie B - matricea tridiagonală asociată unei matrice A

printr-un procedeu oarecare, astfel încît $B=S^{-1}AS$. Dacă notăm elementele b_{ij} ale matricei B prin

$$b_{ii}=a_i \qquad b_{i,i+1}=b_{i+1,i}=b_i \qquad b_{ij}=0 \quad \text{dacă } |i-j|>1 ,$$

Polinomul caracteristic al matricei B este

$$(4.15)\quad \det(B-\lambda I_n)=\begin{vmatrix} a_1-\lambda & b_1 & 0 & 0 & \dots & 0 & 0 \\ b_1 & a_2-\lambda & b_2 & 0 & \dots & 0 & 0 \\ 0 & b_2 & a_3-\lambda & b_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & b_{n-1} & a_n-\lambda \end{vmatrix} = 0 .$$

Presupunem că toţi b_i sînt diferiţi de zero pentru că, în caz contrar, determinantul este produs de doi determinanţi de ordin mai mic. Fie $p_k(\lambda) = \det\begin{pmatrix} 1 & 2 & \dots & k \\ 1 & 2 & \dots & k\end{pmatrix}_{B-\lambda I_n}$. Dezvoltîndu-l după ultima linie obţinem:

$$(4.16)\qquad p_k(\lambda)=(a_k-\lambda)\cdot p_{k-1}(\lambda)-b_{k-2}^2\cdot p_{k-2}(\lambda)$$

$(k=2,3,\dots,n)$. Rădăcinile polinomului $p_k(\lambda)$ sînt valorile proprii ale matricei $\begin{pmatrix} 1 & 2 & \dots & k \\ 1 & 2 & \dots & k\end{pmatrix}_A$ şi sînt reale. Are loc următorul rezultat:

<u>Teorema 4.2 Sirul de polinoame</u> $p_0(\lambda), p_1(\lambda),\dots,p_n(\lambda)$ <u>cu</u> $p_0\equiv 1$, $p_1(\lambda)=a_1-\lambda$, <u>definit prin relaţia de recurenţă</u> (4.16), <u>satisface următoarele condiţii</u>:

1. <u>două funcţii consecutive din şir nu se pot anula simultan</u>,
2. <u>dacă</u> $p_j(c)=0$ <u>pentru</u> $j\in\{1,2,\dots,n-1\}$ <u>, fixat, atunci</u> $p_{j-1}(c)\cdot p_{j+1}(c)<0$,
3. <u>dacă</u> $p_n(c)=0$ <u>atunci</u> $p_n'(c)\cdot p_{n-1}(c)<0$.

<u>Demonstraţie</u> 1. Dacă $p_{k-1}(c)=p_{k-2}(c)=0$ atunci, folosind (4.16) se găseşte că $p_0(c)=0$, lucru imposibil; prin urmare, două funcţii consecutive din şir nu se pot anula simultan.

2. Dacă $p_j(c)=0$, atunci $p_{j+1}(c)=-b_{j-1}^2\cdot p_{j-1}(c)$ şi $p_{j-1}(c)\cdot p_{j+1}(c)=$ $=-b_{j-1}^2\cdot p_{j-1}^2(c)<0$.

2. Derivînd relaţia (4.16) obţinem

$$(4.17)\qquad p_k'(\lambda)=-p_{k-1}(\lambda)+(a_k-\lambda)p_{k-1}'(\lambda)-b_{k-2}^2\cdot p_{k-2}'(\lambda) .$$

Folosind notaţiile

(4.18) $$r_k(\lambda) = p'_{k-1}(\lambda)\cdot p_k(\lambda)-p_{k-1}(\lambda)\cdot p'_k(\lambda)$$

din (4.16) şi (4.17) rezultă

(4.19) $$r_k(\lambda) = p^2_{k-1}(\lambda) + b^2_{k-1}r_{k-1}(\lambda) \quad .$$

Cum $r_2(\lambda) = p_1^2(\lambda)+b_1^2 > 0$, folosind inducţia completă rezultă că $r_k > 0$ pentru $k=2,3,\ldots,n$. Deci, în particular

$$0 < r_n(\lambda)=p'_{n-1}(\lambda)p_n(\lambda)-p'_n(\lambda)p_{n-1}(\lambda) \;.$$

Dacă $p_n(c)=0$, atunci

$$0 < r_n(c) = - p'_n(c)\cdot p_{n-1}(c) \quad .$$

Definiţia 4.1 Fie $I \subset R$ - un interval şi $f:I \longrightarrow R$ o funcţie derivabilă. Se numeşte şir Sturm asociat lui f un şir finit de funcţii continue $f_o, f_1,\ldots,f_s : I \longrightarrow R$ cu proprietăţile:

1°. $f_o \equiv f$,

2°. $f_s(x) \neq 0$, $\forall x \in I$,

3°. dacă există $c \in I$ şi $i \in \{1,2,\ldots,s\}$ astfel încît $f_i(c) = 0$ atunci $f_{i-1}(c)\cdot f_{i+1}(c) < 0$,

4°. dacă există $c \in I$ astfel încît $f(c) = 0$, atunci $f'_o(c)\cdot f_1(c) < 0$.

Observaţia 4.1 Definiţia 4.1 nu coincide cu definiţia 2.2 din Cap.III,B ;cele două definiţii au axiomele 4° diferite. Denumirea de şir Sturm se păstrează şi în acest caz, deoarece şirul construit prin definiţia 4.1 face aceleaşi servicii ca şi un şir Sturm obişnuit şi se mînuieşte la fel de uşor ca acesta. Rezultatul care urmează susţine această afirmaţie.

Teorema 4.3 Dacă $p_o, p_1,\ldots,p_s$ este un şir Sturm pentru funcţia f pe intervalul $[a,b] \subset I$ şi $V(x)$ este numărul schimbărilor de semn în şirul valorilor în x ale celor n+1 funcţii p_i $i=o,1,\ldots,n$, atunci numărul zerourilor reale ale funcţiei f din intervalul (a,b) este $V(b) - V(a)$.

Demonstraţie Se transpune cuvînt cu cuvînt demonstraţia teoremei 2.4 din Cap.III,B cu excepţia ultimei părţi. Mai exact, vom arăta că dacă avem un singur y real (şi numai unul) astfel în-

cît $a<y<b$ şi y este zero pentru o funcţie oarecare p_i, atunci diferenţa $V(b)-V(a)$ este 1 cînd y este zero pentru f şi 0 în caz contrar. Dacă $f(y)=0$, atunci $f(a)f(b)<0$, iar $f'(a)$ şi $f'(b)$ au semne contrare cu p_1. Din teorema lui Lagrange rezultă că $f(x)=(x-y)f'(\eta)$ cu η cuprins între x şi y. Dacă $x\in(a,y)$, atunci $f(x)$ şi $p_1(x)$ au semne contrare, deci $V(b)=V(a)+1$. Dacă y este zero pentru p_i $(i\neq n)$ atunci se arată că $V(a)=V(b)$ la fel ca în demonstraţia Teoremei 2.4 din Cap.IIIB.*

<u>Corolar 4.1</u> <u>Dacă</u> $p_i(\lambda)$ $(i=0,1,2,..,n)$ <u>este şirul de polinoame definit în Teorema 4.2, şirul</u> $f_i=p_{n-i}$ $(i=0,1,2,..,n)$ <u>este şir Sturm asociat polinomului caracteristic</u> $p_n(\lambda)$ <u>al matricei tridiagonale</u> B.*

Ţinînd cont de acest rezultat se poate folosi următorul procedeu de separare a rădăcinilor polinomului $\det(B-\lambda I_n)$.

- Se ia intervalul $(-\|B\|_1, \|B\|_1)$, care conţine toate valorile proprii ale matricei B,

- înjumătăţim acest interval; şirul Sturm pentru fiecare subinterval ne va preciza numărul de rădăcini situate în aceste subintervale; noile intervale se înjumătăţesc din nou şi se determină numărul de rădăcini reale din fiecare subinterval nou (folosind tot şirul Sturm), ş.a.m.d..

Procesul se continuă pînă cînd obţinem subintervale care conţin cîte o singură rădăcină. Folosind apoi metoda bisecţiei se obţin aproximări oricît de bune dorim pentru valorile proprii. Această metodă este foarte stabilă. In plus, valorile proprii multiple se determină la fel cu cele simple.

In momentul în care se cunoaşte valoarea proprie λ, se determină vectorul propriu (al matricei B) corespunzător rezolvînd sistemul cu matrice tridiagonală $(B-\lambda I_n)X=0$, impunînd $x_1=1$. Inductiv se arată că

$$x_i=(-1)^{i-1}p_{i-1}(\lambda)/b_2b_3\dots b_i \quad (i=2,3,..,n).$$

adică vectorii proprii se obţin cu ajutorul polinoamelor şirului Sturm. Cum $B=S^{-1}AS$, dacă u este vector propriu corespunzător valorii proprii λ, rezultă $S^{-1}ASu=Bu=\lambda u$ sau $A(Su)=\lambda(Su)$. Deci Su este vector propriu al matricei A corespunzător lui λ.

§5. Algoritmii LU şi QR

Algoritmul LU

Acest algoritm a fost propus de către H.Rutishauser în anul 1958. El se bazează pe LU-factorizarea unei matrice pătratice, cu $L\in I_1^{(n)}$ şi $U\in S^{(n)}$. Am văzut în §2, Cap.II că o astfel de descompunere există dacă şi numai dacă A are toţi minorii diagonali principali $\neq 0$; în plus, dacă există, descompunerea este unică.

Fie $A_1=A$ - o matrice care are proprietatea K_n. Această matrice se poate scrie în mod unic sub forma:

(5.1) $$A_1 = L_1U_1 \qquad L_1\in I_1^{(n)},\ U_1\in S^{(n)} .$$

Considerăm matricea $A_2=U_1L_1$. In general $A_2\neq A_1$. Dacă A_2 are proprietatea K_n, o putem pune sub forma $A_2=L_2U_2$ cu $L_2\in I_1^{(n)}$ şi $U_2\in S^{(n)}$. Procesul este definit prin recurenţa

(5.2) $$A_k = U_{k-1}L_{k-1} = L_kU_k \qquad k=2,3,\ldots .$$

Cum $A_{k-1}=L_{k-1}U_{k-1}$, rezultă că $A_k=L_{k-1}^{-1}A_{k-1}L_{k-1}$. Inductiv se stabileşte că

$$A_k = L_{k-1}^{-1}L_{k-2}^{-1}\ldots L_2^{-1}L_1^{-1}AL_1L_2\ldots L_{k-2}L_{k-1},$$

sau că

$$L_1L_2\ldots L_{k-1}A_k = AL_1L_2\ldots L_{k-1}.$$

Matricele $T_k=L_1L_2\ldots L_{k-1}$ şi $R_k=U_kU_{k-1}\ldots U_1$ sînt triunghiulară inferior cu 1 pe diagonala principală şi respectiv triunghiulară superior. Atunci au loc egalităţile

$$T_kR_k = L_1L_2\ldots L_{k-1}(L_kU_k)U_{k-1}\ldots U_2U_1 = L_1L_2\ldots L_{k-1}A_kU_{k-1}\ldots U_2U_1 =$$
$$= AL_1L_2\ldots L_{k-1}\cdot U_{k-1}\ldots U_2U_1 = AT_{k-1}R_{k-1} .$$

Cu ajutorul relaţiei $T_kR_k = AT_{k-1}R_{k-1}$ se stabileşte, prin inducţie completă, că

(5.3) $$T_k R_k = A^k .$$

Am obţinut astfel LU-factorizarea matricei A^k.

Teorema 5.1 Fie $A \in \mathcal{M}_n(R)$, cu valori proprii reale şi distincte în modul şi astfel încît algoritmul LU să poată fi aplicat pentru orice k. Dacă matricea P a vectorilor proprii şi matricea $Q=P^{-1}$ admit, fiecare dintre ele LU-factorizare, atunci şirul A_k este convergent la o matrice triunghiulară superior, avînd pe diagonală valorile proprii ale matricei A.

Nu prezentăm demonstraţia acestei teoreme.

R.H.Wilkinson aprecia în anul 1964 că algoritmul LU reprezintă progresul cel mai semnificativ care a fost făcut în legătură cu problema valorilor proprii de la apariţia calculatoarelor electronice.

Algoritmul QR

Acest algoritm a fost dezvoltat de către J.G.F.Francis şi V.N.Kublanovskaja în 1961/62. In locul LU-factorizării se foloseşte QR-factorizarea. In multe situaţii s-a dovedit a fi cea mai eficace metodă pentru rezolvarea problemei valorilor proprii. Un avantaj al utilizării algoritmului QR îl constituie faptul că QR-factorizarea se poate face pentru orice matrice nesingulară, spre deosebire de LU-factorizare care se putea face numai pentru matrice cu proprietatea K_n.

Construcţia algoritmului este similară cu cea a algoritmului LU. Matricea $A_1=A$ se pune sub forma $A_1=Q_1R_1$ cu $Q_1 \in GO_n(R)$ şi $R_1 \in S^{(n)}$ şi, în general, şirul de matrice A_k se obţine prin recurenţa

(5.4) $$A_k = R_{k-1}Q_{k-1} = Q_k A_k \qquad (k=2,3,\dots)$$

cu $Q_k \in GO_n(R)$, $R_k \in S^{(n)}$. Ca la algoritmul LU se arată că

$$Q_1Q_2\dots Q_{k-1}A_k = AQ_1Q_2\dots Q_{k-1} \quad .$$

Cu notaţiile $P_k = Q_1Q_2\dots Q_k$ şi $T_k = R_k\dots R_2R_1$ se obţin egalită-

țile

$$P_kT_k=Q_1Q_2\dots Q_{k-1}(Q_kR_k)R_{k-1}\dots R_2R_1=Q_1Q_2\dots Q_{k-1}A_kR_{k-1}\dots R_2R_1=$$
$$=AP_{k-1}T_k \quad .$$

Folosind inducția completă rezultă

$$P_kT_k = A^k \tag{5.5}$$

unde $P_k \in GO_n(R)$ iar $T_k \in S^{(n)}$. Vom folosi matrice R_k cu diagonala pozitivă (lucru care asigură unicitatea QR-factorizării). In acest caz T_k va avea diagonala pozitivă, iar (5.5) va reprezenta QR-factorizarea unică a matricei A^k.

Tot fără demonstrație prezentăm următorul rezultat

Teorema 5.2 Fie $A \in \mathcal{M}_n(R)$ - o matrice nesingulară avînd valorile proprii λ_i distincte în modul; fie X - matricea vectorilor proprii și $X^{-1}=LU$, LU-factorizarea matricei X^{-1}. Atunci matricele $A_k = a_{ij}^{(k)}$ au proprietățile:

a) $\lim_{k\to\infty} a_{ij}^{(k)} = 0$ pentru $i > j$,

b) $\lim_{k\to\infty} a_{ii}^{(k)} = \lambda_i$ $(i=1,2,\dots,n)$ cu $|\lambda_1|>|\lambda_2|>\dots>|\lambda_n|$,

c) $\lim_{k\to\infty} a_{ij}^{(k)} = t_{ij}$ pentru $i < j$.*

Este remarcabil faptul că concluziile teoremei 5.2 rămîn valabile în condiții mult mai generale (de exemplu, cînd o rădăcină este multiplă iar celelalte sînt de module distincte).

§6. Metoda puterilor

Metoda puterilor

Această metodă se folosește pentru determinarea valorii caracteristice de modul maxim. In particular ea se poate totdeauna folosi pentru matricele care au toate elementele pozitive. Intr-adevăr, pentru astfel de matrice are loc teorema Perron-Frobenius care afirmă:

Dacă $A =[a_{ij}]$ are $a_{ij} > 0$, $i,j=1,2,\dots,n$, atunci există o valoare caracteristică pozitivă a lui A care este dominantă, adică, dacă $\rho(A) = \alpha$, atunci α este valoare caracteristică a lui

A; dacă $\alpha'\neq\alpha$ este tot o valoare caracteristică a lui A atunci $\alpha>|\alpha'|$ şi deci α este rădăcină caracteristică simplă. In plus, vectorul caracteristic al lui A, corespunzător valorii proprii α poate fi ales cu toate componentele pozitive; mai exact, nici-un alt vector caracteristic al matricei A nu poate avea toate componentele nenegative.

Pentru simplificarea prezentării, vom presupune că vectorii proprii $C_i, i=1,2,\dots,n$ generează R^n. Acesta este cazul cînd A este simetrică. Alegem versorul $v^{(o)}$ ($\|v^{(o)}\|_2=1$). Il scriem în baza C_i: $v^{(o)}=\sum_{i=1}^{n}a_iC_i$. Atunci $Av^{(o)}=\sum_{i=1}^{n}a_i\cdot AC_i=\sum_{i=1}^{n}a_i\cdot\lambda_iC_i$ (dacă $AC_i=\lambda_iC_i$). Definim versorul $v^{(1)}$ prin egalitatea

$$Av^{(o)}=\mu^{(1)}v^{(1)},$$

cu $\|v^{(1)}\|_2=1$. In general, definim $v^{(k+1)}$ cu $\|v^{(k+1)}\|_2=1$ prin relaţia de recurenţă

$$Av^{(k)}=\mu^{(k+1)}v^{(k+1)}.$$

Prin inducţie completă se stabileşte că

$$A^kv^{(o)}=\sum_{i=1}^{n}a_i\lambda_i^kC_i=\mu^{(k+1)}v^{(k+1)}.$$

Presupunem că $|\lambda_1|=\rho(A)$. Atunci

$$A^kv^{(o)}=\lambda_1^k\sum_{i=1}^{n}a_i\left(\frac{\lambda_i}{\lambda_1}\right)^kC_i\simeq a_1\lambda_1^kC_1$$

dacă $a_1\neq0$ şi k este suficient de mare. Rezultă că $v^{(k)}\longrightarrow\bar{C}_1$ şi $\mu^{(k)}\longrightarrow\rho(A)=|\lambda_1|$. Condiţia ca $v^{(k)}$ să fie versori (în norma $\|\;\|_2$) nu este esenţială. In practică se obişnuieşte să se normalizeze vectorii luînd prima componentă 1.

Exemplul 6.1 (Todd) Să aplicăm metoda puterilor matricei

$$A=\begin{bmatrix}0,2 & 0,9 & 1,32\\ -11,2 & 22,28 & -10,72\\ -5,8 & 9,45 & -1,94\end{bmatrix}.$$

Luăm $v^{(o)}={}^t[1\quad 0\quad 0]$. Obţinem $Av^{(o)}={}^t[0,2\quad -11,2\quad -5,8]=\mu^{(1)}v^{(1)}$ unde $\mu^{(1)}=0,2$ şi $v^{(1)}={}^t[1\quad -56\quad -29]$. Similar obţinem $A^2v^{(o)}=Av^{(1)}={}^t[-88,48\quad -948\quad -478,7]=\mu^{(2)}v^{(2)}$ unde $\mu^{(2)}=-88,48$ şi $v^{(2)}={}^t[1\quad 10,7143\quad 5,4107]$. Procesul se continuă pînă cînd componentele lui $v^{(k+1)}$ devin „apropiate" de componentele lui $v^{(k)}$,

adică diferențele componentelor celor doi vectori sînt mai mici decît un anumit prag ε ales la început.*

Metoda deflației

Dacă rădăcinile caracteristice ale matricei A satisfac condiția $|\lambda_1|>|\lambda_2|>\ldots>|\lambda_n|$ și am determinat (de exemplu, prin metoda puterilor) pe λ_1 și C_1, apare problema să găsim o matrice A_1 de ordin n-1 pentru care $\lambda_2,\ldots,\lambda_n$ să fie valori proprii, căreia să-i putem aplica metoda puterilor pentru a determina pe λ_2, ș.a.m.d.. Trecerea de la matricea $A\in\mathcal{M}_n(R)$ la matricea $A_1\in\mathcal{M}_{n-1}(R)$ care satisface condiția de mai sus se numește deflație. Vom prezenta în cele ce urmează o metodă de deflație.

Fie $v={}^t[1\ \ \hat{v}]$ un vector caracteristic al matricei A corespunzător valorii proprii (aici $\hat{v}$ notează vectorul (n-1)-dimensional avînd drept componente componentele de pe locurile 2,3,...,n ale vectorului v; vom folosi această convenție și mai departe). Atunci $Av=\lambda v$ se poate scrie sub forma:

$$\begin{bmatrix} a_{11} & {}^t\hat{r} \\ \hat{c} & \hat{A} \end{bmatrix}\begin{bmatrix} 1 \\ \hat{v} \end{bmatrix}=\lambda\begin{bmatrix} 1 \\ \hat{v} \end{bmatrix}, \tag{6.1}$$

care conduce la sistemul

$$\begin{cases} a_{11}+{}^t\hat{r}\cdot\hat{v}=\lambda \\ \hat{c}+\hat{A}\hat{v}=\lambda\hat{v} \end{cases}. \tag{6.2}$$

Dacă

$$S=\begin{bmatrix} 1 & {}^t\hat{0} \\ \hat{v} & I \end{bmatrix} \tag{6.3}$$

rezultă

$$S^{-1}=\begin{bmatrix} 1 & {}^t\hat{0} \\ -\hat{v} & I \end{bmatrix} \qquad S^{-1}AS=\begin{bmatrix} a_{11}+{}^t\hat{r}\cdot\hat{v} & {}^t\hat{r} \\ -a_{11}\hat{v}+\hat{c}-\hat{v}\,{}^t\hat{r}\hat{v}+\hat{A}\hat{v} & \hat{A}-\hat{v}\,{}^t\hat{r} \end{bmatrix}=\begin{bmatrix} \lambda & {}^t\hat{r} \\ \hat{0} & \hat{A}-\hat{v}\,{}^t\hat{r} \end{bmatrix}.$$

Notăm: $A_1=\hat{A}-\hat{v}\,{}^t\hat{r}$. Presupunem că $S^{-1}AS$ are vectorul propriu $w={}^t[1\ \ \hat{w}]$ corespunzător valorii proprii $\mu\neq\lambda$, adică

$$S^{-1}AS\begin{bmatrix} 1 \\ \hat{w} \end{bmatrix}=\mu\begin{bmatrix} 1 \\ \hat{w} \end{bmatrix}. \tag{6.4}$$

Rezultă că

$$\begin{cases} \lambda+{}^t\hat{r}\hat{w}=\mu \\ A_1\hat{w}=\mu\hat{w} \end{cases}.$$

Fie $\tilde{w}$, λ - o pereche caracteristică a lui A_1. Atunci

a) dacă ${}^t\hat{r}\cdot\tilde{w} \neq 0$, vectorul $\hat{w} = \frac{\mu-\lambda}{{}^t\hat{r}\tilde{w}}\tilde{w}$ satisface (6.4) şi ${}^t[1 \quad \hat{w}]$, μ este o pereche caracteristică pentru $S^{-1}AS$ şi

$$\begin{bmatrix} 1 & {}^t\hat{0} \\ \hat{v} & I \end{bmatrix}\begin{bmatrix} 1 \\ (\mu-\lambda)\tilde{w}/({}^t\hat{r}\tilde{w}) \end{bmatrix} = \begin{bmatrix} 1 \\ \hat{v} + (\mu-\lambda)\tilde{w}/({}^t\hat{r}\tilde{w}) \end{bmatrix}$$

este vector caracteristic al lui A corespunzător valorii caracteristice μ;

b) dacă ${}^t\hat{r}\tilde{w} = 0$ atunci ${}^t[0 \quad \tilde{w}]$ este vector caracteristic al lui $S^{-1}AS$ şi de asemenea al lui A, corespunzător lui μ.

Am obţinut deci un proces de deflaţie care conduce de la perechea caracteristică v, λ a matricei $A \in \mathcal{M}_n(R)$ la matricea $A_1 \in \mathcal{M}_{n-1}(R)$ avînd valorile proprii comune cu A.

<u>Metoda iteraţiei inverse (Wielandt)</u>

In general, dacă ştim o valoare caracteristică λ a lui A, vectorul propriu corespunzător v se obţine rezolvînd sistemul $(A-\lambda I)v = 0$. Dar acest procedeu este nesatisfăcător în foarte multe cazuri. Pentru determinarea vectorului v se poate proceda în felul următor: se ia un vector oarecare $v^{(o)}$ şi se determină $v^{(1)}, v^{(2)}, \ldots$ din ecuaţiile

(6.6) $$(A-\tilde{\lambda}I)v^{(r+1)} = v^{(r)}$$

unde $\tilde{\lambda}$ este aproximanta valorii caracteristice λ. Inductiv se stabileşte că

(6.7) $$(A-\tilde{\lambda}I)^r v^{(r)} = v^{(o)}$$.

$\tilde{\lambda}$ nefiind valoare proprie pentru A, matricea $A-\tilde{\lambda}I$ va fi nesingulară şi ca urmare, relaţia (6.7) se poate scrie

(6.8) $$v^{(r)} = ((A-\tilde{\lambda}I)^{-1})^r\, v^{(o)}$$.

Egalitatea evidentă $\det((A-\tilde{\lambda}I)-(\lambda-\tilde{\lambda})I) = \det(A-\lambda I)$ ne spune că, dacă λ este valoare proprie pentru A, atunci $\lambda-\tilde{\lambda}$ este valoare proprie penrt $A-\tilde{\lambda}I$. De aici rezultă că $\frac{1}{(\lambda-\tilde{\lambda})^r}$ este valoare proprie pentru $((A-\tilde{\lambda}I)^{-1})^r$. Dacă vectorii proprii C_i, i=1,2,..n

generează spaţiul R^n atunci

$$v^{(o)} = \sum_{i=1}^{n} a_i C_i$$

şi

$$v^{(r)} = \sum_{i=1}^{n} \frac{1}{(\lambda_i - \tilde{\lambda})^r} a_i C_i$$

deoarece $C_i, \lambda_i - \tilde{\lambda}$ este pereche caracteristică pentru matricea $A - \tilde{\lambda} I$. Presupunînd că $\tilde{\lambda}$ este cel mai aproape de λ_1, termenul $\frac{1}{(\lambda_1 - \lambda)^r} a_1 C_1$ este predominant faţă de ceilalţi termeni; rezultă că $v^{(r)}$ se apropie de C_1 cu atît mai rapid, cu cît $\tilde{\lambda}$ este o aproximaţie mai bună a lui λ_1.

Este remarcabil faptul că aici λ_1 nu mai trebuie să fie de modul maxim. In plus, în vreme ce metoda puterilor implică anumite dificultăţi la implementarea pe calculator, metoda iteraţiei inverse este foarte stabilă la erorile de rotunjire.

§7. Exerciţii

1. Fie

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

a) Calculaţi P^2, P^3, P^4 şi $Q = aI + bP + cP^2 + dP^3$.

b) Care este polinomulcaracteristic al lui P, valorile caracteristice şi vectorii proprii corespunzători?

c) Care sînt valorile caracteristice ale lui Q?

d) Formulaţi rezultatele corespunzătoare pentru matricea

$$A = \begin{bmatrix} a_0 & a_1 & a_2 & \dots & a_n \\ a_n & a_0 & a_1 & \dots & a_{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ a_1 & a_2 & a_3 & \dots & a_0 \end{bmatrix}.$$

e) Discutaţi ca un caz particular pentru d), matricea

$$R = \begin{bmatrix} -2 & 1 & 0 & 1 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 1 & -2 \end{bmatrix}.$$

2. Fie A o matrice pătratică simetrică, cu toţi minorii diagonali principali pozitivi. Arătaţi:

a) $Q = {}^tXAX$ este o formă pătratică pozitiv definită şi reciproc;

b) matricea A are toate valorile proprii pozitive şi reciproc.

3. Desenaţi discurile Gerschgorin ale matricelor

$$A = \begin{bmatrix} 0 & 2 \\ -1 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} 0 & 1 \\ 1 & 3 \end{bmatrix}$$

şi marcaţi pe desene poziţia valorilor caracteristice. Dacă $X = \mathrm{diag}(1, \frac{1}{2})$ calculaţi $A_1 = X^{-1} A X$ şi desenaţi discurile Gerschgorin pentru A_1. Dacă $Y = \mathrm{diag}(1, (3- 5)/2)$ desenaţi discurile Gerschgorin pentru $Y^{-1} B Y$.

4. Găsiţi valoarea caracteristică dominantă şi vectorul caracteristic corespunzător pentru matricea

$$\begin{bmatrix} 133 & 6 & 135 \\ 44 & 5 & 46 \\ -88 & -6 & -90 \end{bmatrix}$$

prin metoda puterilor. Găsiţi de asemenea şi celelalte valori caracteristice şi vectorii proprii corespunzători.

5. (Bodewing). Folosind metoda puterilor determinaţi valoarea caracteristică dominantă a matricei

$$A = \begin{bmatrix} 2 & 1 & 3 & 4 \\ 1 & -3 & 1 & 5 \\ 3 & 1 & 6 & -2 \\ -4 & 5 & -2 & -1 \end{bmatrix}$$

luînd $v^{(0)} = {}^t[1 \ 1 \ 1 \ 1]$.

6. Aplicaţi procedeul iteraţiei inverse (Wielandt) pentru matricea

$$C_4 = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

alegînd $v^{(0)} = {}^t[1 \ 1 \ 1 \ 1]$ şi $\tilde{\lambda} = 0{,}4$.

7. Aplicaţi metodele lui Givens şi Householder pentru a reduce matricea

$$\begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix}$$

la forma tridiagonală, apoi determinaţi valorile caracteristice ale matricei obţinute.

Capitolul V

INTERPOLARE

Interpolarea este una dintre metodele importante de aproximare a funcţiilor. A interpola înseamnă a rezolva o problemă de genul următor: dacă pentru o funcţie de n variabile $f:D \longrightarrow R$ (cu $D \subset R^n$) se cunosc valorile într-un număr de puncte $(x_1^i, x_2^i, \ldots, x_n^i) \in D$, $i=1,2,\ldots,k$, să se determine valoarea ei într-un punct dat $(x_1^o, x_2^o, \ldots, x_n^o) \in D$. Evident, în această formulare generală problema este nedeterminată, căci funcţia f poate avea o valoare arbitrară în punctul dat. De aceea se va presupune că f posedă anumite proprietăţi de continuitate şi diferenţiabilitate.

In prezentul capitol vom avea în vedere numai funcţii reale de o singură variabilă reală. Fie funcţia $f : I \longrightarrow R$, unde $I \subset R$ este un interval şi n+1 puncte aparţinînd lui I: $x_o, x_1, \ldots, x_n$; aceste puncte se numesc <u>noduri de interpolare</u>. Să presupunem că sînt cunoscute valorile funcţiei f în nodurile de interpolare: $y_i = f(x_i)$, $i = o,1,\ldots,n$. Se pune problema construirii unei funcţii g, ce aparţine unei clase de funcţii specificată, care în nodurile de interpolare satisface la anumite condiţii ce o „apropie" de funcţia f. Funcţia g se numeşte funcţia de interpolare a funcţiei f. Cel mai adesea, clasa de funcţii în care se caută g este clasa polinomamelor, a polinoamelor trigonometrice, a funcţiilor spline etc. In privinţa condiţiilor de „apropiere" între f şi g, se cere ca în noduri aceste funcţii să aibă cel puţin valori egale.

Astfel pusă problema, se vede că funcţia poate fi dată numai cu valorile ei în noduri (valori ce ar putea fi obţinute, de exemplu, pe cale experimentală). Valoarea funcţiei f într-un punct $x \in I$ se aproximează cu valoarea funcţiei g în acelaşi punct. Cînd aproximarea se face într-un punct x situat în intervalul cel mai mic ce conţine nodurile $x_0, x_1, \ldots, x_n$ vorbim de interpolare; dacă x se află în afara acestui interval, vorbim de extrapolare.

Clasa polinoamelor este una din clasele de funcţii de interpolare cele mai importante. Aceasta provine din faptul că în urma operaţiilor algebrice cît şi a operaţiilor de derivare şi integrare cu polinoame obţinem ca rezultat tot un polinom. Pe de altă parte, o funcţie continuă poate fi aproximată uniform pe un interval compact cu un polinom, căci are loc următoarea

Teoremă de aproximare a lui Wierstrass. Dacă funcţia $f: [a,b] \longrightarrow R$ este continuă, atunci pentru orice $\varepsilon > 0$ există un polinom P_ε astfel încît

$$|f(x) - P_\varepsilon(x)| < \varepsilon \quad , \quad (\forall)\ x \in [a,b] .$$

In [6, pag.215-224] sînt date patru demonstraţii pentru această teoremă remarcabilă. Totuşi acest rezultat nu poate fi utilizat în calcule numerice, căci este garantată existenţa polinomului de aproximare uniformă dar nu şi faptul că acest polinom poate fi construit cu ajutorul unui algoritm.

Formula lui Taylor poate fi privită ca o formulă de interpolare pe un singur nod. Intr-adevăr, dacă $f \in C^{k+1}(I)$, pentru orice $x \in I$ există un punct c cuprins între x şi x_0 încît

$$f(x) = T_k(x) + \frac{(x - x_0)^{k+1}}{(k+1)!} f^{(k+1)}(c) .$$

Condiţiile de apropiere între funcţia f şi polinomul T_k sînt concentrate în punctul x_0:

$$\left[T_k(x)\right]^{(m)}_{x=x_o} = f^{(m)}(x_o) \quad , \qquad m = o,1,2,\dots,k$$

Utilizarea polinomului Taylor T_k pentru aproximarea funcţiei f este limitată la punctele x suficient de aproape de x_o, căci eroarea $|f(x) - T_k(x)|$ este suficient de mică numai pentru astfel de puncte. Pe de altă parte, în realitate polinomul de interpolare (sau funcţia de interpolare) trebuie să satisfacă condiţii de apropiere de funcţia dată într-un număr mai mare de noduri.

§ 1. Formula de interpolare a lui Lagrange

Fie $f:I \longrightarrow R$ o funcţie definită pe intervalul $I \subset R$ şi nodurile de interpolare $x_o,x_1,\dots,x_n$ (presupunem că ele sînt puncte distincte ale intervalului I).

__Teorema 1.1. Există un polinom unic P de grad cel mult n şi care în nodurile de interpolare satisface condiţiile__

(1.1) $$P(x_i) = f(x_i) \quad , \quad i = o,1,\dots,n \ .$$

__Demonstraţie.__ Să arătăm mai întîi că pentru orice $i=o,1,\dots$ $\dots,n$ există un polinom L_i de grad cel mult n şi care satisface condiţiile mai simple

(1.2) $$L_i(x_j) = \delta_{ij} \ , \qquad j = o,1,\dots,n,$$

unde δ_{ij} este simbolul lui Kronecker. Intr-adevăr, din relaţiile (1.2) rezultă că nodurile $x_o,x_1,\dots,x_{i-1},x_{i+1},\dots,x_n$ sînt rădăcini ale polinomului L_i şi ca urmare acesta are forma

(1.3) $$L_i(x) = c_i(x-x_o)(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n) \ .$$

Pentru a determina constanta c_i, punem condiţia $L_i(x_i) = 1$, obţinută din (1.2) pentru $j = i$; avem

$$1 = c_i(x_i-x_o)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i - x_n) \ ,$$

de unde

$$c_i = \frac{1}{(x_i-x_o)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i - x_n)}$$

Inlocuind în (1.3), ajungem la forma definitivă a polinomului L_i:

(1.4) $$L_i(x)=\frac{(x-x_o)(x-x_1)\ \dots\ (x-x_{i-1})(x-x_{i+1})\ \dots\ (x-x_n)}{(x_i-x_o)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}$$

Existența unui polinom P de grad cel mult n și care în noduri verifică condițiile (1.1) se arată acum imediat. Intr-adevăr, un astfel de polinom poate fi polinomul P dat de

(1.5) $$P(x) = \sum_{i=o}^{n} f(x_i)L_i(x) \quad .$$

Pentru a dovedi unicitatea acestui polinom să presupunem că ar exista două polinoame P și Q de grad cel mult n și care în noduri satisfac la condițiile (1.1). Atunci polinomul R = P - Q este de asemenea de grad cel mult n și se anulează în $x_o,x_1,\dots,x_n$. Deoarece aceste n+1 zerouri ale lui R au fost presupuse distincte, rezultă că R = o, adică P și Q coincid. Q.e.d.

<u>Definiție</u>. Polinomul a cărui existență și unicitate este dată de teorema precedentă se numește <u>polinomul de interpolare</u> al funcției f pe nodurile $x_o,x_1,\dots,x_n$.

<u>Observație</u>. Dacă nu impunem condiția ca gradul polinomului P(x) să fie cel mult n, se pot determina o infinitate de polinoame ce au valoarea $f(x_i)$ în punctul x_i, pentru orice i=o,1,...,n. De exemplu, polinomul

$$P(x) + (x-x_o)(x-x_1)\dots(x-x_n)\ Q(x)\ ,$$

unde P(x) este polinomul de interpolare pe nodurile $x_o,x_1,\dots,x_n$ iar Q(x) este un polinom nenul arbitrar, verifică condițiile (1.1), dar are gradul cel puțin n+1.

Polinomul de interpolare (care este unic !) poate fi scris în forme diferite, utile atît din punct de vedere teoretic cît

şi practic. Forma acestui polinom, indicată de (1.5), se numeşte forma lui Lagrange a polinomului de interpolare sau, pe scurt, polinomul lui Lagrange. Vom nota cu $L(x_0,x_1,\dots,x_n;\ f|x)$ polinomul de interpolare asociat funcţiei f pe nodurile $x_0,x_1,\dots,x_n$ sub forma lui Lagrange. Aşadar,

$$(1.6)\qquad L(x_0,x_1,\dots,x_n;\ f|x) = \sum_{i=0}^{n} f(x_i)\,\frac{(x-x_0)(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})(x-x_{i+2})\dots(x-x_n)}{(x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}.$$

Pentru a-i da o formă mai simplă, introducem polinomul $\omega(x)$ prin

$$(1.7)\qquad \omega(x) = \prod_{i=0}^{n}(x-x_i) = (x-x_0)(x-x_1)\dots(x-x_n)$$

Evident,

$$\omega'(x_i) = (x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)$$

şi, ca urmare, (1.6) devine

$$(1.8)\qquad L(x_0,x_1,\dots,x_n;\ f|x) = \sum_{i=0}^{n} f(x_i)\,\frac{\omega(x)}{(x-x_i)\,\omega'(x_i)}\ .$$

Formula (1.8) este incomod de utilizat atunci cînd trebuie aproximate valorile funcţiei date prin valorile polinomului lui Lagrange în mai multe puncte x ale intervalului I. Intr-adevăr, din relaţia (1.7) se vede că trebuie refăcut calculul lui $\omega(x)$ pentru fiecare valoare a lui x în parte. Acest fapt comportă un număr mare de operaţii. Această dificultate este înlăturată prin punerea polinomului lui Lagrange sub aşa-numita formă baricentrică. In acest scop, să observăm că pentru $f(x) = P_k(x)$, unde P_k este un polinom de grad $k \le n$, avem $L(x_0,x_1,\dots,x_n;\ f|x) = P_k(x)$. In cazul particular $f(x) = 1$, această egalitate se scrie

$$\sum_{i=0}^{n} \frac{\omega(x)}{(x-x_i)\,\omega'(x_i)} = 1$$

Ca urmare,

$$L(x_0,x_1,\dots,x_n;\ f|x) = \sum_{i=0}^{n} f(x_i)\frac{\omega(x)}{(x-x_i)\omega'(x_i)} =$$

$$= \frac{\displaystyle\sum_{i=0}^{n} f(x_i)\frac{\omega(x)}{(x-x_i)\,\omega'(x_i)}}{\displaystyle\sum_{i=0}^{n}\frac{\omega(x)}{(x-x_i)\,\omega'(x_i)}}$$

şi, simplificînd $\omega(x)$,

$$(1.9)\qquad L(x_0,x_1,\dots,x_n;\ f|x) = \frac{\displaystyle\sum_{i=0}^{n} f(x_i)\frac{1}{(x-x_i)\,\omega'(x_i)}}{\displaystyle\sum_{i=0}^{n}\frac{1}{(x-x_i)\omega'(x_i)}}\ ,$$

care este forma baricentrică a polinomului lui Lagrange. Avantajul acestei forme provine din faptul că nu mai conţine $\omega(x)$ iar numerele $\omega'(x_i)$ se calculează o singură dată (nu depind de x). Ca urmare, este asigurat un grad mai ridicat de precizie şi o rapiditate de calcul mai mare.

Să definim funcţie $R_n: I \longrightarrow R$ prin

$$(1.10)\qquad R_n(x) = f(x) - L(x_0,x_1,\dots,x_n;\ f|x)\ ,$$

adică eroarea cu care f(x) este aproximată de valoarea polinomului lui Lagrange în x. Această funcţie se numeşte <u>restul de ordin</u> <u>n</u> asociat funcţiei f pe nodurile $x_1,x_2,\dots,x_n$. Din (1.10) decurge

$$(1.11)\qquad f(x) = L(x_0,x_1,\dots,x_n;\ f|x) + R_n(x),\qquad x\in I\ ,$$

care este <u>formula de interpolare a lui Lagrange</u> de ordin n.

Se mai scrie, avînd în vedere (1.8) şi (1.6),

(1.12) $$f(x) = \sum_{i=0}^{n} f(x_i) \frac{\omega(x)}{(x-x_i)\,\omega'(x_i)} + R_n(x), \quad x \in I ,$$

sau

$$f(x) = \sum_{i=0}^{n} f(x_i) \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} +$$

$$+ R_n(x), \qquad x \in I$$

Vom pune în evidenţă cîteva cazuri particulare ale formulei de interpolare.

a) Cazul $n = 1$ (interpolarea lineară). Avem două noduri: x_0 şi x_1. Dacă notăm $y_0 = f(x_0)$ şi $y_1 = f(x_1)$, formula de interpolare a lui Lagrange, de ordinul 1, se scrie:

$$f(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 + R_1(x) \quad , \quad x \in I$$

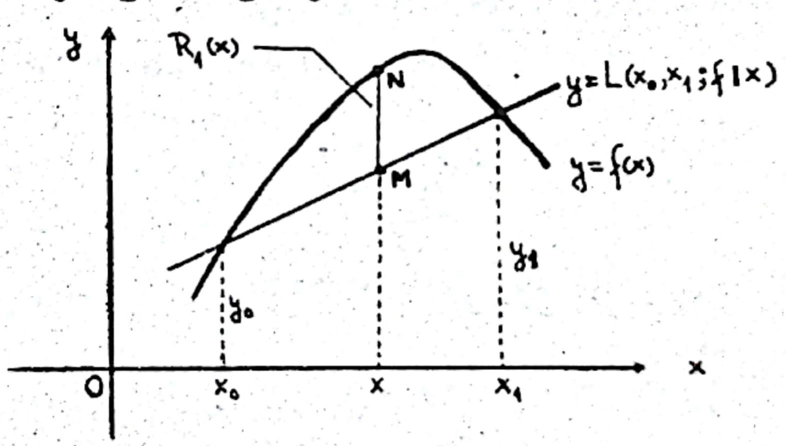


Fig.1.1.

Geometric, graficul polinomului lui Lagrange este o dreaptă ce trece prin punctele (x_0,y_0) şi (x_1,y_1). Aproximarea funcţiei f cu polinomul lui Lagrange revine la înlocuirea graficului acestei funcţii cu porţiunea dreptei ce corespunde la $x \in I$. Restul în punctul x, $R_1(x)$, are semnificaţia de lungime a segmentului orientat MN.

b. Cazul n = 2 (interpolarea parabolică). Nodurile de interpolare sînt: x_o, x_1, x_2. Fie $y_i = f(x_i)$, i = o,1,2, Polinomul lui Lagrange este de gradul doi şi are drept grafic o parabolă cu axa paralelă la Oy şi trecînd prin punctele (x_o, y_o), (x_1, y_1) şi (x_2, y_2). In acest caz, formula de interpolare se scrie:

$$f(x) = \frac{(x-x_1)\ (x-x_2)}{(x_o-x_1)(x_o-x_2)}\, y_o + \frac{(x-x_o)\ (x-x_2)}{(x_1-x_o)(x_1-x_2)}\, y_1 + \frac{(x-x_o)\ (x-x_1)}{(x_2-x_o)(x_2-x_1)}\, y_2 + R_2(x),\ x \in I\ .$$

Exemplul 1.1. O funcţie f fiind dată prin tabelul de valori

x	1,0	1,3	1,6	1,9
f(x)	0,7651977	0,6200860	0,4554022	0,2818186

(funcţia Bessel $J_o(x)$), să se aproximeze valoarea ei în punctul x = 1,5 utilizînd polinomul lui Lagrange de gradul 3.

Avem, considerînd $x_o = 1{,}o$, $x_1 = 1{,}3$, $x_2 = 1{,}6$, $x_4 = 1{,}9$:

$$f(1{,}5) \simeq \frac{(1{,}5-1{,}3)(1{,}5-1{,}6)(1{,}5-1{,}9)}{(1-1{,}3)(1-1{,}6)(1-1{,}9)} \cdot 0{,}7651977 +$$

$$+ \frac{(1{,}5-1)(1{,}5-1{,}6)(1{,}5-1{,}9)}{(1{,}3-1)(1{,}3-1{,}6)(1{,}3-1{,}9)} \cdot 0{,}6200860 +$$

$$+ \frac{(1{,}5-1)(1{,}5-1{,}3)(1{,}5-1{,}9)}{(1{,}6-1)(1{,}6-1{,}3)(1{,}6-1{,}9)} \cdot 0{,}4554022 + \frac{(1{,}5-1)(1{,}5-1{,}3)(1{,}5-1{,}6)}{(1{,}9-1)(1{,}9-1{,}3)(1{,}9-1{,}6)} \cdot 0{,}2818186$$

şi, în urma calculelor, $f(1{,}5) \simeq 0{,}5118200$.

Pentru a putea evalua eroarea aproximării funcţiei f cu polinomul de interpolare, vom preciza forma restului R_n. Dacă funcţia f satisface condiţii suplimentare, este posibil să dăm restului R_n o expresie similară cu forma lui Lagrange pentru restul formulei lui Taylor.

Teorema 1.2. Dacă $f \in C^{n+1}(I)$ şi $x \in I$, există un punct c situat în intervalul cel mai mic ce conţine punctele $x, x_o, x_1, \ldots$ $\ldots, x_n$ astfel încît

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-x_o)(x-x_1)\ldots(x-x_n) \quad (1.13)$$

Demonstraţie. Să definim funcţia ajutătoare $\varphi : I \to R$ prin

$$\varphi(z) = R_n(z) - k(z-x_o)(z-x_1)\ldots(z-x_n) \quad , \quad z \in I \, , \quad (1.14)$$

unde constanta reală k este deocamdată arbitrară. Formula (1.14) se mai scrie

$$\varphi(z) = f(z) - L(x_o, x_1, \ldots, x_n; f|z) - k(z-x_o)\ldots(z-x_n), \quad z \in I. \quad (1.15)$$

Este evident că $\varphi \in C^{n+1}(I)$ şi că nodurile $x_o, x_1, \ldots, x_n$ sînt zerouri ale funcţiei .

Fie $t \in I$ definit de $x_o, x_1, \ldots, x_n$. Să determinăm k astfel încît punctul t să fie un zerou al funcţiei φ. Ţinînd seama de (1.14) aceasta are loc pentru

$$k = \frac{R_n(t)}{(t-x_o)(t-x_1)\ldots(t-x_n)} \, . \quad (1.16)$$

Dacă în (1.14) şi (1.15) considerăm k dat de (1.16), funcţia φ are n+2 zerouri distincte: $t, x_o, x_1, \ldots, x_n$. In acest caz, funcţia φ' are, conform teoremei lui Rolle, n+1 zerouri distincte situate în cel mai mic interval ce conţine punctele $t, x_o, \ldots, x_n$. Aplicînd succesiv aceeaşi teoremă funcţiilor $\varphi', \varphi'', \ldots, \varphi^{(n)}$, sîntem conduşi de faptul că există un punct c aparţinînd intervalului menţionat astfel încît

$$\varphi^{(n+1)}(c) = o$$

In virtutea relaţiei (1.15), această egalitate revine la

$$f^{(n+1)}(c) - k(n+1)! = o \, ,$$

de unde, avînd în vedere faptul că numărul k este dat de (1.16), obţinem

$$R_n(t) = \frac{f^{(n+1)}(c)}{(n+1)!}(t-x_0)(t-x_1)\dots(t-x_n)\ . \tag{1.17}$$

Cum relaţia (1.17) este adevărată şi în cazul cînd t ar coincide cu un oricare dintre aceste noduri, conchidem că pentru orice $t \in I$ există un punct c aşa ca (1.17) are loc. Teorema este complet demonstrată.

<u>Observaţie</u>. Dacă $f^{(n+1)}$ este mărginită pe I şi notăm

$$M = \sup_{x\in I} \left| f^{(n+1)}(x) \right| ,$$

obţinem următoarea limitare pentru restul formulei de interpolare a lui Lagrange:

$$|f(x) - L(x_0,x_1,\dots,x_n;\ f|x)| \le \frac{M}{(n+1)!}|(x-x_0)\dots(x-x_n)|,\ x\in I. \tag{1.18}$$

Evident, utilitatea practică a relaţiei precedente este destul de restrînsă. Foarte adesea funcţia se dă printr-un tabel al valorilor ei într-un număr de puncte şi, ca urmare, nu se ştie nimic despre derivatele ei. In practică se calculează valorile diverselor polinoame ale lui Lagrange pînă se obţine o aproximaţie convenabilă în punctul dat.

Vom stabili o formulă de recurenţă care va permite calcularea valorii polinomului lui Lagrange într-un punct, fără a calcula coeficienţii lui.

<u>Teorema 1.3. Polinomul lui Lagrange</u> $L(x_0,\dots,x_n;\ f|x)$ <u>satisface la următoarea relaţie de recurenţă:</u>

$$L(x_0,x_1,\dots x_n;\ f|x) = \tag{1.19}$$

$$= \frac{(x-x_0)L(x_1,x_2,\dots,x_n;\ f|x)-(x-x_n)L(x_0,x_1,\dots,x_{n-1};f|x)}{x_n - x_0}\ .$$

<u>Demonstrație</u>. Să notăm cu P polinomul de grad cel mult n din membrul doi al egalității (1.19). In conformitate cu Teorema 1.1 polinomul de interpolare este unic. De aceea, este suficient să arătăm că au loc relațiile $P(x_i)= f(x_i)$, $i=0,1,\dots,n$, pentru ca P să coincidă cu polinomul lui Lagrange. Intr-adevăr, avem

$$P(x_o)=\frac{-(x_o-x_n)L(x_o,x_1,\dots,x_{n-1};\ f|x_o)}{x_n-x_o}=f(x_o)\ ,$$

$$P(x_n)=\frac{(x_n-x_o)L(x_1,x_2,\dots,x_n;\ f|x_n)}{x_n-x_o}=f(x_n)\ ,$$

$$P(x_i)=\frac{(x_i-x_o)L(x_1,\dots,x_n;\ f|x_i)-(x_i-x_n)L(x_o,\dots,x_{n-1};f|x_i)}{x_n-x_o}=$$

$$=\frac{(x_i-x_o)f(x_i)-(x_i-x_n)f(x_i)}{x_n-x_o}=f(x_i),\ i=1,2,\dots,n-1,$$

ceea ce încheie demonstrația.

<u>Observație</u>. Formula (1.19) se poate pune sub forma

$$L(x_o,x_1,\dots,x_n;\ f|x)=\frac{\begin{vmatrix} x-x_o & L(x_o,x_1,\dots,x_{n-1};\ f|x) \\ x-x_n & L(x_1,x_2,\dots,x_n\ ;\ f\,|\,x)\end{vmatrix}}{x_n-x_o} \tag{1.20}$$

Această formulă oferă o regulă comodă pentru construirea din aproape în aproape a următorului tabel al valorilor polinoamelor lui Lagrange într-un punct x dat:

x_o	$x-x_o$	$L(x_o;f\|x)$				
x_1	$x-x_1$	$L(x_1;f\|x)$	$L(x_o,x_1;f\|x)$			
x_2	$x-x_2$	$L(x_2;f\|x)$	$L(x_1,x_2;f\|x)$	$L(x_o,x_1,x_2;f\|x)$		
x_3	$x-x_3$	$L(-_3;f\|x)$	$L(x_2,x_3;f\|x)$	$L(x_1,x_2,x_3;f\|x)$	$L(x_o,x_1,x_2,x_3;f\|x)$	

- -

Dacă ultima valoare aproximativă $L(x_o,x_1,x_2,x_3;\ f|x)$ nu-i suficient de bună se ia un nou nod, x_4, şi se formează linia corespunzătoare lui în tabelul de mai sus. Se compară $L(x_o,\ldots$ $\ldots,x_4;\ f|x)$ şi $L(x_o,\ldots,x_3;\ f|x)$ pentru a vedea dacă aproximarea este suficient de bună.

Acest mod de calculare a valorilor polinoamelor lui Lagrange într-un punct dat se numeşte metoda Aitken - Neville. Pentru degajarea unui algoritm, să facem notaţia:

$$L_{ij}(x) = L(x_{i-j},x_{i-j+1},\ldots,x_i;\ f|x)\ , \quad j \le i\ .$$

Conform formulei (1.19) avem:

$$L_{ij}(x)= \frac{(x-x_i)\ L_{i-1,j-1}(x)-(x-x_{i-j}).L_{i,j-1}(x)}{x_{i-j} - x_i}$$

Tabelul de mai sus se scrie (convenim ca $L_{io}= f(x_i)$, pentru orice i):

x_o	L_{oo}			
x_1	L_{1o}	L_{11}		
x_2	L_{2o}	L_{21}	L_{22}	
x_3	L_{3o}	L_{31}	L_{32}	L_{33}
- - -	- - -	- - -	- - -	- - -

<u>Algoritmul Aitken-Neville de interpolare</u>. Pentru a aproxima valoarea funcţiei f în punctul x cunoscînd valorile ei în nodurile $x_o,x_1,x_2,\ldots,$ procedăm astfel:

<u>Pasul 1</u>. Pune $L_{oo} = f(x_o)$

<u>Pasul 2</u>. Pune $i = 1$

<u>Pasul 3</u>. Pune $L_{io} = f(x_i)$

<u>Pasul 4</u>. Calculează $L_{ij}= \dfrac{(x-x_i)L_{i-1,j-1}-(x-x_{i-j})L_{i,j-1}}{x_{i-j} - x_i}$, $j=1,2,\ldots,i$

<u>Pasul 5</u>. Dacă L_{ii} este o aproximaţie suficient de bună, treci la Pasul 7. In caz contrar treci la Pasul 6.

Pasul 6. Adunăm 1 la i şi treci la Pasul 3.

Pasul 7. Consideră $f(x) = L_{ii}$. Procedura este completă.

Pentru a decide dacă L_{ii} este o aproximaţie suficient de bună se poate folosi condiţia

$$|L_{ii} - L_{i-1,i-1}| < \varepsilon ,$$

unde ε este o toleranţă dată pentru eroare.

§ 2. Formula de interpolare a lui Hermite

In cele ce urmează va fi generalizată o parte din consideraţiile din paragraful precedent.

Fie funcţia $f:I \to R$, n+1 noduri distincte, $x_o, x_1, \dots, x_n$, situate în intervalul I şi numerele naturale nenule $m_o, m_1, \dots, m_n$ - ordinele de multiplicitate ale nodurilor. Presupunem că $f \in C^{m-1}(I)$, unde $m = \max\{m_o, m_1, \dots, m_n\}$. Fie $p=(m_o+m_1+\dots+m_n)-1$.

Teorema 2.1. Există un polinom unic H de grad cel mult p care satisface în noduri la condiţiile:

$$H(x_o)=f(x_o), H'(x_o)=f'(x_o), H''(x_o)=f''(x_o), \dots, H^{(m_o-1)}(x_o)=f^{(m_o-1)}(x_o);$$

$$H(x_1)=f(x_1), H'(x_1)=f'(x_1), H''(x_1)=f''(x_1), \dots, H^{(m_1-1)}(x_1)=f^{(m_1-1)}(x_1);$$

- -

$$H(x_n)=f(x_n), H'(x_n)=f'(x_n), H''(x_n)=f''(x_n), \dots, H^{(m_n-1)}(x_n)=f^{(m_n-1)}(x_n).$$

Demonstraţie. Să construim mai întîi, pentru orice $i = o,1,\dots,n$ şi orice $j = o,1,\dots,m_i-1$, un polinom H_{ij} de grad cel mult p şi care verifică condiţiile:

$$H_{ij}(x_o)=o,\ H'_{ij}(x_o)=o, \dots, H^{(j)}_{ij}(x_o)=o, \dots, H^{(m_o-1)}_{ij}(x_o)=o ;$$

- -

$$H_{ij}(x_i)=o,\ H'_{ij}(x_i)=o, \dots, \boxed{H^{(j)}_{ij}(x_i)=1}, \dots, H^{(m_i-1)}_{ij}(x_i)=o; \tag{2.2}$$

- -

$$H_{ij}(x_n)=o,\ H'_{ij}(x_n)=o, \dots, H^{(j)}_{ij}(x_n)=o, \dots, H^{(m_n-1)}_{ij}(x_n)=o ,$$

adică

$$H_{ij}^{(k)}(x_\ell)=\delta_{i\ell}\,\delta_{jk},\quad \ell=0,1,\dots,n\ ,\ k=0,1,\dots,m_i-1$$

Cum din (2.2) rezultă că $x_0,x_1,\dots,x_n$ sînt zerouri ale polinomului H_{ij} de ordine de multiplicitate cel puţin $m_0,m_1,\dots,m_{i-1},j,m_{i+1},\dots,m_n$, respectiv, rezultă că acest polinom este de forma:

$$(2.3)\qquad H_{ij}(x)=(x-x_0)^{m_0}(x-x_1)^{m_1}\dots$$
$$\dots(x-x_{i-1})^{m_{i-1}}(x-x_i)^j(x-x_{i+1})^{m_{i+1}}\dots(x-x_n)^{m_n}Q_{ij}$$

sau, scris concentrat,

$$(2.4)\qquad H_{ij}(x)=\frac{\omega(x)}{(x-x_i)^{m_i}}(x-x_i)^j\,Q_{ij}(x)\ ,$$

unde

$$(2.5)\qquad \omega(x)=\prod_{i=0}^{n}(x-x_i)^{m_i}$$

iar Q_{ij} este un polinom deocamdată nedeterminat. Din (2.4) rezultă că grad $Q_{ij}\le p$ - grad $\omega - j + m_i = m_i-j-1$. Din faptul că x_i este rădăcină a lui H_{ij} de ordin de multiplicitate j, urmează că Q_{ij} nu are ca rădăcină pe x_i.

Pentru a determina polinomul Q_{ij}, să dezvoltăm H_{ij} după formula lui Taylor în jurul punctului x_i. In virtutea condiţiilor (2.2) avem:

$$H_{ij}(x)=\frac{1}{j!}(x-x_i)^j+c(x-x_i)^{m_i}+\dots\ ,$$

de unde

$$\frac{H_{ij}(x)}{(x-x_i)^j}=\frac{Q_{ij}(x)\,\omega(x)}{(x-x_i)^{m_i}}=\frac{1}{j!}+c(x-x_i)^{m_i-j}+\dots$$

şi deci,

$$Q_{ij}(x)=\frac{(x-x_i)^{m_i}}{j!\,\omega(x)}+\frac{1}{\frac{\omega(x)}{(x-x_i)^{m_i}}}\left[c(x-x_i)^{m_i-j}+\dots\right]\ .$$

Deoarece polinomul $\frac{\omega(x)}{(x-x_i)^{m_i}}$ are o dezvoltare în jurul punctului $x=x_i$ de forma

$$d_0 + d_1(x-x_i)+\ldots \quad , \quad d_0 \neq 0 ,$$

rezultă că

$$Q_{ij}(x)= \frac{(x-x_i)^{m_i}}{j!\,\omega(x)} + (x-x_i)^{m_i-j} \frac{c+\ldots}{d_0+d_1(x-x_i)+\ldots}$$

De aici şi din faptul că grad $Q_{ij} \leq m_i - j - 1$, urmează că Q_{ij} este suma termenilor de grad cel mult $m_i - j - 1$ din dezvoltarea Taylor a fracţiei $\frac{(x-x_i)^{m_i}}{j!\,\omega(x)}$ în jurul punctului $x = x_i$. In consecinţă, obţinem:

$$Q_{ij}(x) = \frac{1}{j!} \sum_{k=0}^{m_i-j-1} \frac{1}{k!} (x-x_i)^k \left[\frac{(x-x_i)^{m_i}}{\omega(x)}\right]^{(k)}_{x=x_i} .$$

Inlocuind această expresie a lui Q_{ij} în (2.4), ajungem la următoarea formă pentru polinomul H_{ij} căutat:

$$(2.6) \qquad H_{ij}(x)= \frac{1}{j!} \sum_{k=0}^{m_i-j-1} \frac{1}{k!} \left[\frac{(x-x_i)^{m_i}}{\omega(x)}\right]^{(k)}_{x=x_i} \frac{\omega(x)}{(x-x_i)^{m_i-j-k}} .$$

Acum este simplu de construit un polinom care să răspundă la cerinţele teoremei. Intr-adevăr, polinomul H dat de

$$(2.7) \qquad H(x) = \sum_{i=0}^{n} \sum_{j=0}^{m_i-1} f^{(j)}(x_i).H_{ij}(x)$$

este de grad cel mult p şi satisface condiţiile (2.1).

Unicitatea unui astfel de polinom se demonstrează prin reducere la absurd. Presupunem că ar exista două astfel de polinoame, H şi G. Atunci polinomul R = H - G este de grad cel mult

p şi are rădăcinile $x_o, x_1, \ldots, x_n$ de ordine de multiplicitate $m_o, m_1, \ldots, m_n$, respectiv. Ca urmare R este de grad cel mult p şi are $\sum_{o}^{n} m_i = p+1$ rădăcini, ceea ce este absurd. Teorema este complet demonstrată.

Polinomul construit în teorema precedentă se numeşte <u>polinomul de interpolare al lui Hermite</u> pentru funcţia f pe nodurile $x_o, x_1, \ldots, x_n$ de ordine de multiplicitate $m_o, m_1, \ldots, m_n$, respectiv. Să-l notăm cu $H(x_o, m_o; x_1, m_1; \ldots; x_n, m_n; f|x)$. Din (2.6) şi (2.7) rezultă că acest polinom are forma:

(2.8) $$H(x_o, m_o; x_1 m_1; \ldots; x_n, m_n; f|x) =$$

$$= \sum_{i=o}^{n} \sum_{j=o}^{m_i-1} \sum_{k=o}^{m_i-j-1} \frac{1}{j!\,k!} f^{(j)}(x_i) \left[\frac{(x-x_i)^{m_i}}{\omega(x)}\right]^{(k)}_{x=x_i} \frac{\omega(x)}{(x-x_i)^{m_i-j-k}}$$

unde $\omega(x)$ este dat de (2.5).

<u>Observaţii</u>. 1) Dacă nu cerem polinomului lui Hermite să fie de grad cel mult p, unicitatea lui nu mai este asigurată. Intr-adevăr, un polinom de forma

$$H(x_o, m_o; \ldots; x_n, m_n; f|x) + Q(x).\ \omega(x)$$

satisface condiţiile (2.1), dar este de grad mai mare ca p pentru orice polinom Q nenul.

2) Polinomul lui Hermite generealizează în acelaşi timp şi polinomul Taylor şi polinomul lui Lagrange. Pentru n = o (2.8) devine:

$$H(x_o, m_o; f|x) = \sum_{j=o}^{m_o-1} \frac{(x-x_o)^j}{j!} f^{(j)}(x_o) \; ,$$

care este polinomul Taylor de grad m_o-1 asociat funcţiei f în punctul x_o.

Pentru $m_0 = m_1 = \ldots = m_n = 1$ se obţine

$$H(x_0,1;\ x_1,1;\ldots;x_n,1;f|x) = \sum_{i=0}^{n} f(x_i)\left[\frac{x-x_i}{\omega(x)}\right]_{x=x_i} \frac{\omega(x)}{x-x_i} =$$

$$= \sum_{i=0}^{n} f(x_i)\,\frac{\omega(x)}{(x-x_i)\,\omega'(x_i)}$$

unde $\omega(x) = \prod_{i=0}^{n}(x-x_i)$, adică tocmai polinomul lui Lagrange $L(x_0,x_1,\ldots,x_n;\ f|x)$.

Să definim funcţia $R_p : I \to R$, funcţia rest de ordin p, prin

(2.9) $\quad R_p(x) = f(x) - H(x_0,m_0;\ldots;x_n,m_n;\ f|x)\ ,\quad x\in I$

Ca urmare putem scrie:

(2.10) $\quad f(x) = H(x_0,m_0;x_1,m_1;\ldots;x_n,m_n;f|x) + R_p(x)\ ,\quad x\in I\ .$

Aceasta este <u>formula de interpolare a lui Hermite</u> pentru funcţia f şi nodurile $x_0,x_1,\ldots,x_n$ multiple de ordin $m_0,m_1,\ldots,m_n$, respectiv.

Teorema următoare dă o formă pentru restul R_p al acestei formule şi este o generalizare a Teoremei 1.2. Omitem demonstraţia ei, căci urmează pas cu pas pe cea dată Teoremei 1.2. In [18, p.30] se poate găsi această demonstraţie în detaliu.

<u>Teorema 2.2. Dacă $f\in C^{p+1}(I)$ şi $x\in I$ este un punct arbitrar atunci există c situat în intervalul cel mai mic ce conţine punctele $x,x_0,x_1,\ldots,x_n$ astfel încît</u>

(2.11) $\quad R_p(x) = \dfrac{f^{(p+1)}(c)}{(p+1)!}\,(x-x_0)^{m_0}(x-x_1)^{m_1}\ldots(x-x_n)^{m_n}\ .$

Dacă $f^{(p+1)}$ este mărginită pe I şi $M = \sup_{x\in I}|f^{(p+1)}(x)|$, formula (2.11) ne conduce la următoarea evaluare a erorii formulei de interpolare a lui Hermite.

$$(2.12) \qquad |f(x)-H(x_o,m_o;\ldots;x_n,m_n;f|x)| \le \frac{M}{(p+1)!}|\omega(x)|, \quad x\in I,$$

ω avînd expresia (2.5).

Un caz particular important este <u>formula de interpolare a lui Hermite pe noduri duble</u> ($m_o= m_1=\ldots= m_n= 2$). Graficele funcţiei date f şi polinomului $H(x_o,2;\ldots;x_n,2;\ f|x)$ se intersectează în puncte de abscise $x_o,x_1,\ldots,x_n$ şi au tangente comune în aceste puncte. Să arătăm că polinomul lui Hermite se poate pune în acest caz sub forma:

$$(2.13) \qquad H(x_o,2;\ldots;x_n,2;f|x)=\sum_{i=o}^{n}\left[1-2\,L_i'(x_i)(x-x_i)\right]\left[L_i(x)\right]^2 f(x_i)+$$

$$+\sum_{i=o}^{n}(x-x_i)\left[L_i(x)\right]^2 f'(x_i)\ ,$$

unde

$$(2.14) \qquad L_i(x)=\frac{(x-x_o)(x-x_1)\ \ldots\ (x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_o)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}\ ,$$

$$i=o,1,\ldots,n$$

(a se vedea şi (1.4)).

Intr-adevăr, din formula (2.8) decurge că

$$(2.15) \qquad H(x_o,2;\ldots;x_n,2;f|x)=\sum_{i=o}^{n}\left\{f(x_i)\left[\frac{(x-x_i)^2}{\omega(x)}\right]_{x=x_i}\frac{\omega(x)}{(x-x_i)^2}+\right.$$

$$\left.+\ f(x_i)\left[\frac{(x-x_i)^2}{\omega(x)}\right]'_{x=x_i}\frac{\omega(x)}{x-x_i}\right\}+\sum_{i=o}^{n}f'(x_i)\left[\frac{(x-x_i)^2}{\omega(x)}\right]_{x=x_i}\frac{\omega(x)}{x-x_i}$$

unde $\omega(x) = (x-x_o)^2(x-x_1)^2\ldots(x-x_n)^2$. Cum însă,

$$\left[\frac{(x-x_i)^2}{\omega(x)}\right]'_{x=x_i}\frac{\omega(x)}{x-x_i}=$$

$$=\left[\frac{1}{\left[L_i(x)\right]^2\left[(x_i-x_o)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)\right]^2}\right]'_{x=x_i}\times$$

$$\times [L_i(x)]^2 (x-x_i) [(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)]^2 =$$

$$= \left[\frac{1}{[L_i(x)]^2}\right]'_{x=x_i} \cdot [L_i(x)]^2 (x-x_i) = \frac{-2\, L_i'(x_i)}{[L_i(x_i)]^3} [L_i(x)]^2 (x-x_i) =$$

$$= -2\, L_i'(x_i) [L_i(x)]^2 (x-x_i)$$

şi, în urma unor calcule asemănătoare,

$$\left[\frac{(x-x_i)^2}{\omega(x)}\right]_{x=x_i} \frac{\omega'(x)}{(x-x_i)^2} = [L_i(x)]^2,$$

$$\left[\frac{(x-x_i)^2}{\omega(x)}\right]_{x=x_i} \frac{\omega(x)}{x-x_i} = [L_i(x)]^2 (x-x_i),$$

prin înlocuire în (2.15) se obţine formula (2.13).

§ 3. Diferenţe finite şi divizate

In prezentul paragraf va fi pus la punct instrumentul de lucru necesar pentru dezvoltările ce vor urma. Operatorii liniari ce vor fi definiţi acţionează pe mulţimea funcţiilor reale.

Fie $h \in R_+$ un număr numit pas şi $f : R \to R$ o funcţie arbitrară. Se numeşte operator de translaţie de pas h, operatorul E_h definit prin

(3.1) $$(E_h f)(x) = f(x+h), \quad x \in R$$

Se numeşte operator de diferenţă progresivă (sau diferenţă ascendentă) de pas h, operatorul Δ_h definit prin

(3.2) $$(\Delta_h f)(x) = f(x+h) - f(x), \quad x \in R$$

In mod similar, se numeşte operator de diferenţă regresivă (sau diferenţă descendentă) de pas h, operatorul ∇_h definit prin

(3.3) $$(\nabla_h f)(x) = f(x) - f(x-h), \quad x \in R .$$

In sfîrşit, se numeşte <u>operator de diferenţă centrală</u> de pas h, operatorul δ_h definit prin

(3.4) $$(\delta_h f)(x) = f(x + \frac{h}{2}) - f(x - \frac{h}{2}) , \quad x \in R .$$

Vom renunţa la scrierea indicelui h atunci cînd acest fapt nu va crea confuzii.

Puterile acestor operatori se introduc prin recurenţă:

(3.5) $$E^{n+1} = E(E^n), \quad \Delta^{n+1} = \Delta(\Delta^n), \quad \nabla^{n+1} = \nabla(\nabla^n), \quad \delta^{n+1} = \delta(\delta^n), \quad n \quad N ,$$

(s-a subînţeles că $E^1 = E$, $\Delta^1 = \Delta$ etc.). Puterea Δ^n se mai numeşte <u>operator de diferenţă progresivă de ordin n</u>. Denumiri simi- pentru puterile celorlalţi operatori. Se notează cu I <u>operatorul</u> identic $(I f)(x) = f(x)$, $x \in R$ şi se face convenţia utilă: $E^o = \Delta^o = \nabla^o = \delta^o = I$.

Se verifică imediat că operatorul de translaţie cît şi cele trei tipuri de operatori de diferenţă de mai sus sînt liniari. De asemenea are loc

$$T^m(T^n f) = T^{m+n} f , \quad m,n \in N ,$$

unde T este unul dintre aceşti operatori.

<u>Exemplul 3.1</u>. Să se arate că $\Delta^{n+1} P_n(x) = o$, unde $P_n(x)$ este un polinom de grad cel mult n. Acelaşi rezultat, dacă operatorul Δ este înlocuit cu ∇ sau δ.

Intr-adevăr, ţinînd seama de liniaritatea lui Δ şi faptul uşor de verificat că $\Delta(x^k)$ este un polinom de grad k-1, rezultă că $\Delta P_n(x)$ este un polinom de grad cel mult n-1. In final, $\Delta^n P_n(x)$ este o constantă şi deci $\Delta^{n+1} P_n(x) = o$.

<u>Teorema 3.1</u>. <u>Au loc următoarele formule</u>:

(3.6) $$f(x + nh) = \sum_{k=o}^{n} C_n^k \Delta^k f(x) ,$$

şi

(3.7) $$\Delta^n f(x) = \sum_{k=o}^{n} (-1)^k C_n^k f[x + (n-k)h] \quad .$$

Demonstraţia acestor egalităţi se face fără dificultăţi prin inducţie completă şi o omitem.

Observaţie. Utilizarea operatorului de translaţie E ar permite scrierea relaţiilor (3.6) şi (3.7) funcţional. Sub forma dată sînt foarte sugestive: prima exprimă valorile succesive ale funcţiei f prin diferenţele sale de diverse ordine şi a doua exprimă diferenţa de ordin n a funcţiei f prin valorile sale succesive. Aceste relaţii pot fi scrise simbolic astfel:

$$E^n = (I + \Delta)^n \text{ şi } \qquad \Delta^n = (E - I)^n \quad .$$

In ambele formule intră în joc reţeaua de numere reale $x + n\,h$, $n \in N$, determinată de x şi pasul h. Evident, punctele acestei reţele sînt echidistante. In general, utilizarea operatorilor E, Δ, ∇, δ presupune astfel de reţele de puncte.

O schemă cu diferenţe pe reţeaua echidistantă (de pas h) $x_k = x_o + k\,h$, $k = o,1,\ldots,5$, se prezintă sub forma următorului tabel:

x	f(x)	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$	$\Delta^5 f(x)$
x_o	$f(x_o)$					
		$\Delta f(x_o)$				
x_1	$f(x_1)$		$\Delta^2 f(x_o)$			
		$\Delta f(x_1)$		$\Delta^3 f(x_o)$		
x_2	$f(x_2)$		$\Delta^2 f(x_1)$		$\Delta^4 f(x_o)$	
		$\Delta f(x_2)$		$\Delta^3 f(x_1)$		$\Delta^5 f(x_o)$
x_3	$f(x_3)$		$\Delta^2 f(x_2)$		$\Delta^4 f(x_1)$	
		$\Delta f(x_3)$		$\Delta^3 f(x_2)$		
x_4	$f(x_4)$		$\Delta^2 f(x_3)$			
		$\Delta f(x_4)$				
x_5	$f(x_5)$					

Vedem că numerele $\Delta^k f(x_o)$, $k = o,1,\ldots,5$, se află pe linia înclinată în jos spre dreapta. Pe de altă parte, deoarece

(3.8) $$\Delta = E\nabla \quad \text{şi} \quad \nabla = E^{-1}\Delta$$

avem că: $\nabla f(x_5) = \Delta f(x_4)$, $\nabla^2 f(x_5) = \Delta^2 f(x_3)$, $\nabla^3 f(x_5) = \Delta^3 f(x_2)$, $\nabla^4 f(x_5) = \Delta^4 f(x_1)$ şi $\nabla^5 f(x_5) = \Delta^5 f(x_o)$. Observăm că $\nabla^k f(x_5)$, $k = o,1,\ldots,5$, se află pe linia înclinată în sus spre dreapta. Mai avem

(3.9) $$\Delta = E^{\frac{1}{2}}\delta \quad \text{şi} \quad \delta = E^{-\frac{1}{2}}\Delta \quad ;$$

ca urmare, avem de exemplu: $\delta^2 f(x_2) = E^{-1}\Delta^2 f(x_2) = \Delta^2 f(x_1)$, $\delta^4 f(x_2) = E^{-2}\Delta^4 f(x_2) = \Delta^4 f(x_o)$ etc. Numerele $\delta^{2k} f(x_2)$ $k=o,1,2$, se găsesc pe o linie orizontală. Se desprinde faptul că schema cu diferenţe este aceeaşi şi că este doar o chestiune de notaţie cum sînt numite diferenţele. De exemplu,

$$\Delta^3 f(x_1) = \nabla^3 f(x_4) = \delta^3 f(x_{\frac{5}{2}}) \quad .$$

Exemplul 3.2. Să se construiască schema cu diferenţe pentru funcţia

$$f(x) = x^4 + x^2 + 1$$

plecînd de la $x_o = o$, cu pasul $h = 1$.

x	f(x)	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$	$\Delta^5 f(x)$
0	1					
		2				
1	3		16			
		18		36		
2	21		52		24	
		70		60		0
3	91		112		24	
		182		84		
4	273		196			
		378				
5	651					

Teorema 3.2. Dacă $f \in C^n[x, x+nh]$, atunci există un număr real θ, $0 < \theta < 1$, astfel încît

$$(3.10) \qquad \Delta^n f(x) = h^n . f^{(n)}(x + \theta\, n\, h)$$

Demonstrație. Prin inducție completă. Pentru $n = 1$, formula (3.10) se reduce la formula creșterilor finite. Presupunem că pentru $k < n$ avem

$$\Delta^k f(x) = h^k f^{(k)}(x + \theta' k\, h) \quad , \quad 0 < \theta' < 1 \quad .$$

Atunci

$$\Delta^{k+1} f(x) = \Delta^k [f(x+h) - f(x)] = h^k [f^{(k)}(x+h+ \theta' k\, h) - f^{(k)}(x + \theta' k\, h)]$$

și, ținînd seama de formula creșterilor finite,

$$\Delta^{k+1} f(x) = h^{k+1} f^{(k+1)}(x + \theta' k\, h + \theta'' h) \; ,$$

cu $\theta'' \in (0,1)$. Fie

$$\theta = \frac{\theta' k + \theta''}{k + 1}$$

și să observăm că avem $0 < \theta < 1$. Cum avem

$$\Delta^{k+1} f(x) = h^{k+1} f^{(k+1)}(x + \theta(k+1)h), \qquad 0 < \theta < 1 \; ,$$

rezultă afirmația teoremei.

Observație. Formula (3.10) se scrie

$$f^{(n)}(x + \theta\, n\, h) = \frac{\Delta^n f(x)}{h^n}$$

și, deoarece $f^{(n)}$ este continuă, avem

$$(3.11) \qquad f^{(n)}(x) = \lim_{h \to 0} \frac{\Delta^n f(x)}{h^n} \quad .$$

Ca urmare, obținem formula aproximativă

$$f^{(n)}(x) \simeq \frac{\Delta^n f(x)}{h^n}$$

şi faptul că eroarea tinde la zero cînd $h \to 0$.

x

x x

Am văzut că atunci cînd se construieşte o schemă cu diferenţe finite sînt cunoscute valorile unei funcţii în punctele unei reţele echidistante. Situaţia care apare cel mai adesea în practică este aceea cînd valorile funcţiei sînt cunoscute în punctele unei reţele arbitrare. In acest caz, locul diferenţelor finite este luat de aşa-numitele diferenţe divizate.

Fie $x_0, x_1, \ldots, x_n, \ldots$ un şir de valori distincte ale variabilei x a funcţiei f. Se numeşte diferenţă divizată de ordinul întîi a funcţiei f relativ la x_i şi x_j, $i,j \in N$, $i \neq j$, expresia $f[x_i, x_j]$ definită prin

$$f[x_i, x_j] = \frac{f(x_j) - f(x_i)}{x_j - x_i} \quad (3.12)$$

Diferenţele divizate de ordin superior se introduc prin recurenţă. Se numeşte diferenţă divizată de ordinul $k \geq 2$ a funcţiei f relativ la $x_i, x_{i+1}, \ldots, x_{i+k}$, $i \in N$, expresia $f[x_i, x_{i+1}, \ldots, x_{i+k}]$ definită prin

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i} \quad (3.13)$$

Se face convenţia ca diferenţa de ordin zero a funcţiei f relativă la x_i să fie egală cu $f(x_i)$, adică $f[x_i] = f(x_i)$.

O schemă cu diferenţe divizate se prezintă sub formă de tabel:

x	$f(x)$	1	2	3	4	5
x_0	$f(x_0)$					
		$f[x_0,x_1]$				
x_1	$f(x_1)$		$f[x_0,x_1,x_2]$			
		$f[x_1,x_2]$		$f[x_0,x_1,x_2,x_3]$		
x_2	$f(x_2)$		$f[x_1,x_2,x_3]$		$f[x_0,x_1,x_2,x_3,x_4]$	
		$f[x_2,x_3]$		$f[x_1,x_2,x_3,x_4]$		$f[x_0,x_1,x_2,x_3,x_4,x_5]$
x_3	$f(x_3)$		$f[x_2,x_3,x_4]$		$f[x_1,x_2,x_3,x_4,x_5]$	
		$f[x_3,x_4]$		$f[x_2,x_3,x_4,x_5]$		
x_4	$f(x_4)$		$f[x_3,x_4,x_5]$			
		$f[x_4,x_5]$				
x_5	$f(x_5)$					

<u>Exemplul 3.3</u>. Schema cu diferențe divizate pentru funcția dată de

x	-1	0	0,5	1,3	2,1
$f(x)$	-1	1	1,625	4,497	12,361

este următoarea:

x	$f(x)$	1	2	3	4
-1	-1				
		2			
0	1		-0,5		
		1,25		1	
0,5	1,625		1,8		0
		3,59		1	
1,3	4,497		3,9		
		9,83			
2,1	12,361				

<u>Exemplul 3.4</u>. Diferența divizată de ordin n+1 a unui polinom de grad cel mult n este nulă pentru orice sistem de n+2 puncte distincte între ele.

Intr-adevăr, se poate proceda ca în Exemplul 3.1.

<u>Teorema 3.3</u>. <u>Diferența divizată de ordinul n a funcției f are forma</u>:

(3.14) $$f[x_0,x_1,\dots,x_n]=$$

$$=\sum_{k=0}^{n}\frac{f(x_k)}{(x_k-x_0)(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}\ .$$

<u>Demonstrație</u>. Procedăm prin inducție completă. Pentru n=0 formula (3.14) este verificată căci revine la $f[x_0]=f(x_0)$. Pentru $n=1$ avem:

$$f[x_0,x_1]=\frac{f[x_1]-f[x_0]}{x_1-x_0}=\frac{f(x_0)}{x_0-x_1}+\frac{f(x_1)}{x_1-x_0}$$

și din nou (3.14) are loc. Dacă $n=3$, avem:

$$f[x_o,x_1,x_2]=\frac{f[x_1,x_2]-f[x_o,x_1]}{x_2-x_o}=$$

$$=\frac{1}{x_2-x_o}\left(\frac{f(x_1)}{x_1-x_2}+\frac{f(x_2)}{x_2-x_1}-\frac{f(x_o)}{x_o-x_1}-\frac{f(x_1)}{x_1-x_o}\right)=$$

$$=\frac{f(x_o)}{(x_o-x_1)(x_o-x_2)}+\frac{f(x_1)}{(x_1-x_o)(x_1-x_2)}+\frac{f(x_2)}{(x_2-x_o)(x_2-x_1)}.$$

In mod asemănător, presupunînd că diferenţele divizate de ordin k au forma dată de (3.14), se arată că şi cele de ordin k+1 se pot scrie în această formă. Demonstraţia este încheiată.

<u>Observaţie</u>. Cu ajutorul relaţiei (3.14) se poate vedea imediat că diferenţa divizată de ordin n $f[x_o,x_1,\ldots,x_n]$ are proprietăţile: 1) este liniară în raport cu funcţia, 2) este simetrică în raport cu argumentele $x_o,x_1,\ldots,x_n$, 3) este continuă în raport cu variabilele $x_o,x_1,\ldots,x_n$.

Legătura între diferenţele divizate şi diferenţele finite este precizată de

<u>Teorema 3.4. Fie $f : R \longrightarrow R$, $x\in R$ şi $h > o$. Pentru orice $n \in N$ are loc egalitatea:</u>

$$(3.15)\qquad f[x,x+h,\ldots,x+n\cdot h]=\frac{\Delta^n f(x)}{n!\;h^n}$$

<u>Demonstraţie</u>. Relaţia (3.14) pentru $x_k = x + k\,h$, $k=o,1,\ldots,n$, devine:

$$f[x,x+h,\ldots,x+n\cdot h]=\sum_{k=o}^{n}\frac{f(x+k\cdot h)}{(-1)^{n-k}h^n\,k!\,(n-k)!}=$$

$$=\frac{1}{n!\;h^n}\sum_{k=o}^{n}(-1)^{n-k}C_n^{n-k}f(x+kh)=\frac{1}{n!\;h^n}\,\Delta^n f(x)\;,$$

ultimul egal avînd loc în virtutea relaţiei (3.7).

Se poate da un înţeles diferenţelor divizate şi în cazul cînd două sau mai multe argumente coincid. Astfel, considerăm că

$$f[x_o,x_o] = f'(x_o)$$

motivat de faptul că

$$f[x_o,x_o] = \lim_{h\to o} f[x_o,x_o+h] = \lim_{h\to o} \frac{f(x_o+h)-f(x_o)}{h} = f'(x_o) .$$

In mod analog, definim

$$(3.16)\qquad f[\underbrace{x_o,x_o,\ldots,x_o}_{k+1 \text{ argumente}}] = \frac{1}{k!} f^{(k)}(x_o)$$

motivat prin faptul că, în virtutea formulelor (3.15) şi (3.11), avem

$$f[\underbrace{x_o,x_o,\ldots,x_o}_{k+1}] = \lim_{h\to o} f[x_o,x_o+h,\ldots,x_o+kh] = \lim_{h\to o} \frac{\Delta^k f(x_o)}{k!\ h^k} = \frac{1}{k!} f^{(k)}(x_o)$$

De asemenea,

$$f[x_o,x_1,\ldots,x_n,x,x] = \frac{d}{dx} f[x_o,x_1,\ldots,x_n,x],$$

căci

$$f[x_o,x_1,\ldots,x_n,x,x] = \lim_{h\to o} f[x_o,x_1,\ldots,x+h,x] =$$

$$= \lim_{h\to o} \frac{1}{h}\ f[x_o,x_1,\ldots,x_n,x+h] - f[x_o,x_1,\ldots,x_n,x] =$$

$$= \frac{d}{dx} f[x_o,x_1,\ldots,x_n,x] .$$

In general, dacă $f\in C^{n+k+1}(I)$ şi $x\in I$, atunci

$$(3.17)\qquad f[x_o,x_1,\ldots,x_n,\underbrace{x,x,\ldots,x}_{k+1 \text{ ori}}] = \frac{1}{k!} \frac{d^k}{dx^k} f[x_o,x_1,\ldots,x_n,x],\ k\in N .$$

§ 4. Formula de interpolare a lui Newton. Formule de de interpolare pe noduri echidistante

Diferenţele divizate constituie un instrument cu mari posibilităţi în interpolarea pe noduri non-echidistante. Pentru polinomul de interpolare, a cărui existenţă şi unicitate este garantată de Teorema 1.1, vom da o formă diferită de cea a lui Lagrange.

Fie date o funcţie $f : I \longrightarrow R$ şi $n+1$ noduri de interpolare, $x_o, x_1, \ldots, x_n$, situate în intervalul I. In conformitate cu formulele (3.12) şi (3.13) prin care au fost definite diferenţele divizate, putem scrie egalităţile:

$$
(4.1)\qquad
\begin{aligned}
f(x) &= f[x_o] + (x-x_o)\, f[x,x_o],\\
f[x,x_o] &= f[x_o,x_1] + (x-x_1)\, f[x,x_o,x_1],\\
f[x,x_o,x_1] &= f[x_o,x_1,x_2] + (x-x_2)\, f[x,x_o,x_1,x_2],\\
&- -\\
f[x,x_o,\ldots,x_{n-1}] &= f[x_o,x_1,\ldots,x_n] + (x-x_n) f[x,x_o,x_1,\ldots,x_n].
\end{aligned}
$$

Inmulţind a doua egalitate cu $(x-x_o)$, a treia cu $(x-x_o)(x-x_1)$ ş.a.m.d. şi în sfîrşit ultima cu $(x-x_o)(x-x_1)\ldots$ $\ldots(x-x_{n-1})$ şi adunînd obţinem:

$$
(4.2)\qquad f(x) = f[x_o] + (x-x_o) f[x_o,x_1] + (x-x_o)(x-x_1) f[x_o,x_1,x_2] + \ldots
$$
$$
\ldots + (x-x_o)(x-x_1)\ldots(x-x_{n-1}) f[x_o,x_1,\ldots,x_n] + R_n(x),
$$

unde

$$
(4.3)\qquad R_n(x) = f[x,x_o,x_1,\ldots,x_n] \prod_{i=o}^{n} (x-x_i) \quad , \qquad x \in I .
$$

Egalitatea (4.2) este de forma $f(x) = P(x) + R_n(x)$, unde $P(x)$ este un polinom de grad n (cel mult). Să arătăm că $P(x)$ este tocmai polinomul de interpolare pe nodurile $x_o, x_1, \ldots, x_n$. Aceasta revine la a verifica că $P(x_i) = f(x_i)$, $i=o,1,\ldots,n$. Evident, $P(x_o) = f(x_o)$. Apoi

$$
P(x_1) = f[x_o] + (x_1 - x_o)\, f[x_o,x_1]
$$

şi, în virtutea primei relaţii (4.1) scrise pentru $x = x_1$, urmează că $P(x_1) = f(x_1)$. Mai departe,

$$P(x_2) = f[x_o] + (x_2-x_o)\, f[x_o,x_1] + (x_2-x_o)(x_2-x_1)\, f[x_o,x_1,x_2]$$

şi, ţinînd seama de primele două relaţii (4.1) scrise pentru $x = x_2$, obţinem $P(x_2) = f(x_2)$, etc.

Aşadar, relaţia (4.2) este o formulă de interpolare. Aceasta este formula de interpolare a lui Newton cu diferenţe divizate. Evident, polinomul $P(x)$ este identic cu polinomul lui Lagrange $L(x_o,x_1,\dots,x_n;\, f|x)$ şi restul $R_n(x)$ dat de (4.3) coincide cu cel din formula de interpolare a lui Lagrange pe nodurile $x_o,x_1,\dots,x_n$. Dacă $f \in C^{n+1}(I)$, conform Teoremei 1.2 avem

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} \prod_{i=o}^{n} (x-x_i) \quad ,$$

unde c aparţine intervalului determinat de $x,x_o,x_1,\dots,x_n$. De aici şi din (4.3), rezultă că

$$f[x,x_o,x_1,\dots,x_n] = \frac{f^{(n+1)}(c)}{(n+1)!} \quad .$$

Altfel spus, dacă $f \in C^n(I)$, există un punct c situat în intervalul determinat de $x_o,x_1,\dots,x_n$ astfel încît

(4.4) $$f[x_o,x_1,\dots,x_n] = \frac{f^{(n)}(c)}{n!} \quad .$$

La limită, această formulă devine (3.16).

Exemplul 4.1. Să se găsească polinomul de interpolare satisfăcînd datele: (-1, -1), (o,1), (o,5; 1,625), (1,3; 4,497), (2,1; 12, 361).

Schema cu diferenţe divizate corespunzătoare acestor date a fost construită în Exemplul 3.3. Diferenţele divizate de care avem nevoie în scrierea polinomului de interpolare se găsesc pe

diagonala descendentă spre dreapta (vezi schema 1). Avem:

$$P_4(x) = -1+(x+1).2+(x+1)(x-o).(-o,5)+(x+1)(x-o)(x-o,5).1+$$
$$+(x+1)(x-o)(x-o,5r)(x-1,3).o$$

sau, în urma efectuării calculelor,

$$P_4(x) = x^3 + x + 1.$$

<u>Observație</u>. Din formula (4.2) rezultă că diferența divizată $f[x_o,x_1,\ldots,x_n]$ este coeficientul lui x^n, adică este coeficientul termenului de grad n al polinomului de interpolare pe nodurile $x_o,x_1,\ldots,x_n$. Acest fapt se ia adesea ca definiție pentru diferențele divizate. In acest caz, relația de recurență de tip (3.13) - care a fost pusă la baza definiției diferențelor divizate - decurge din relația de recurență (1.19) pentru polinomul de interpolare pe $x_o,x_1,\ldots,x_n$.

x

x x

Cazurile cînd nodurile de interpolare sînt echidistante prezintă un interes practic deosebit.

Fie $h > o$ pasul rețelei de noduri, adică

(4.5) $$x_i = x_o + ih \ , \quad i = o,1,2,\ldots,n$$

Să efectuăm schimbarea de variabilă

(4.6) $$x = x_o + th$$

în formula de interpolare a lui Newton(4.2). Deoarece $x-x_i=(t-i)h$, $x_i-x_j=(i-j)h$, $(i,j = o,1,\ldots,n)$ și ținînd seama de (3.15), obținem

$$f(x_o+th) = f(x_o) + th\,\frac{1}{1!h}\,\Delta f(x_o) + t(t-1)h^2.\,\frac{1}{2!\,h^2}\,\Delta^2 f(x_o) +$$
$$\ldots + t(t-1)\ldots(t-n+1)h^n.\,\frac{1}{n!\,h^n}\,\Delta^n f(x_o) + R_n$$

sau

$$(4.7)\quad f(x_o+th)=f(x_o)+\frac{t}{1!}\Delta f(x_o)+\frac{t(t-1)}{2!}\Delta^2 f(x_o)+\ldots$$
$$\ldots+\frac{t(t-1)\ldots(t-n+1)}{n!}\Delta^n f(x_o)+R_n$$

Aceasta este <u>formula lui Newton progresivă</u> (sau <u>ascendentă</u>) pe noduri echidistante. Ea poate fi folosită dacă $o < t < 1$, adică punctul x este în apropierea punctului x_o. Dacă x este mai aproape de nodul x_n, atunci $f(x_n)$ este o primă aproximare pentru $f(x)$ şi aproximări mai bune se obţin utilizînd valorile funcţiei f în celelalte noduri: x_{n-1}, x_{n-2}, etc. Să efectuăm schimbarea de variabilă

$$(4.8)\qquad x = x_n + th$$

în formula (4.2). Ţinînd seama că în acest caz $x-x_i=(x_n-x_o)+(t-i)h = (t+n-i)h$, obţinem :

$$f(x_n+th)=f(x_n)+th\frac{1}{1!h}\Delta f(x_{n-1})+t(t+1)h^2\frac{1}{2!\,h^2}\Delta^2 f(x_{n-2})+\ldots$$
$$\ldots+t(t+1)\ldots(t+n-1)h^n\frac{1}{n!\,h^n}\Delta^n f(x_o)+R_n$$

sau

$$(4.9)\quad f(x_n+th)=f(x_n)+\frac{t}{1!}\Delta f(x_{n-1})+\frac{t(t+1)}{2!}\Delta^2 f(x_{n-2})+\ldots$$
$$\ldots+\frac{t(t+1)\ldots(t+n-1)}{n!}\Delta^n f(x_o)+R_n\ ,$$

care este <u>formula lui Newton regresivă</u> (sau <u>descendentă</u>). Este utilă pentru $-1 < t < o$. Deoarece $\Delta^k f(x_{n-k}) = \nabla^k f(x_n)$, $k=o,1,\ldots,n$, avem următoarea formă mai simplă:

$$(4.10)\quad f(x_n+th)=f(x_n)+\frac{t}{1!}\nabla f(x_n)+\frac{t(t+1)}{2!}\nabla^2 f(x_n)+\ldots$$
$$\ldots+\frac{t(t+1)\ldots(t+n-1)}{n!}\nabla^n f(x_n)+R_n\ .$$

In schema cu diferenţe finite (vezi schema din § 3 !), diferenţele ce apar în formula (4.7) se află pe diagonala de sus iar cele din (4.9) sau (4.10) pe diagonala de jos.

Cu o utilizare mai largă sînt acele formule de interpolare care fac apel la diferenţele centrale, adică acelea ce antrenează valorile funcţiei în noduri situate de ambele părţi ale punctului dat x. Aşadar, să presupunem că punctul x se află în apropierea mijlocului intervalului nodurilor. Vom nota cu x_o nodul cel mai apropiat de x şi vom considera că avem 2n+1 noduri de interpolare:

$$(4.11) \qquad x_i = x_o + ih \quad , \qquad i = -n, -(n-1), \ldots, -1, o, 1, \ldots, n \ .$$

Să aplicăm formula de interpolare a lui Newton (4.2) în cazul cînd nodurile (4.11) sînt ordonate astfel:

$$(4.12) \qquad x_o, x_1, x_{-1}, x_2, x_{-2}, \ldots, x_n, x_{-n};$$

obţinem:

$$f(x) = f[x_o] + (x-x_o)\, f[x_o,x_1] + (x-x_o)(x-x_1) f[x_o,x_1,x_{-1}] +$$
$$+ (x-x_o)(x-x_1)(x-x_{-1}) f[x_o,x_1,x_{-1},x_2] + \ldots$$
$$\ldots + (x-x_o)(x-x_1)(x-x_{-1})\ldots(x-x_{n-1})(x-x_{-(n-1)}) f[x_o,x_1,x_{-1},\ldots,x_n] +$$
$$+ (x-x_o)(x-x_1)(x-x_{-1})\ldots(x-x_n) f[x_o,x_1,x_{-1},\ldots,x_n,x_{-n}] + R_{2n+1}(x)$$

sau, efectuînd schimbarea de variabilă (4.6).

$$f(x_o+th) = f(x_o) + \frac{t}{1!}\Delta f(x_o) + \frac{t(t-1)}{2!}\Delta^2 f(x_{-1}) + \frac{t(t^2-1)}{3!}\Delta^3 f(x_{-1}) +$$
$$\ldots + \frac{t(t^2-1)\ldots(t^2-(n-1)^2)}{(2n-1)!}\Delta^{2n-1} f(x_{-(n-1)}) +$$
$$+ \frac{t(t^2-1)\ldots(t^2-(n-1)^2)(t-n)}{(2n)!}\Delta^{2n} f(x_{-n}) + R_{2n+1}$$

Cum $\Delta = E^{\frac{1}{2}}\delta$, formula precedentă se pune sub forma mai simplă

$$(4.13) \quad f(x_o+th) = f(x_o) + \binom{t}{1}\delta f(x_{\frac{1}{2}}) + \binom{t}{2}\delta^2 f(x_o) + \binom{t+1}{3}\delta^3 f(x_{\frac{1}{2}}) +$$
$$+ \binom{t+1}{4}\delta^4 f(x_o) + \ldots + \binom{t+n-1}{2n-1}\delta^{2n-1} f(x_{\frac{1}{2}}) + \binom{t+n-1}{2n}\delta^{2n} f(x_o) + R_{2n+1},$$

unde s-a convenit ca

$$\binom{s}{k} = \frac{s(s-1)\dots(s-k+1)}{k!} \quad .$$

Formula (4.13) este <u>formula de interpolare a lui Gauss progresivă</u>. In mod analog, dacă în loc de (4.12) considerăm următoarea ordonare a nodurilor de interpolare:

(4.14) $$x_0, x_{-1}, x_1, x_{-2}, x_2, \dots, x_{-n}, x_n \;,$$

obţinem

(4.15) $$f(x_0+th) = f(x_0) + \binom{t}{1}\delta f(x_{-\frac{1}{2}}) + \binom{t+1}{2}\delta^2 f(x_0) + \binom{t+1}{3}\delta^3 f(x_{-\frac{1}{2}}) +$$

$$+\binom{t+2}{4}\delta^4 f(x_0) + \dots + \binom{t+n-1}{2n-1}\delta^{2n-1} f(x_{-\frac{1}{2}}) + \binom{t+n}{2n}\delta^{2n} f(x_0) + R_{2n+1};$$

care este <u>formula de interpolare a lui Gauss regresivă</u>.

In [7] pot fi găsite formulele de interpolare ale lui Sterling şi Bessel.

In practica utilizării formulelor (4.13) şi (4.15) se alcătuieşte mai întîi schema diferenţelor centrale. Pentru n = 2, ea are forma:

x	$f(x)$	δ	δ^2	δ^3	δ^4
x_{-2}	$f(x_{-2})$				
		$\delta f(x_{-\frac{3}{2}})$			
x_{-1}	$f(x_{-1})$		$\delta^2 f(x_{-1})$		
		$\delta f(x_{-\frac{1}{2}})$		$\delta^3 f(x_{-\frac{1}{2}})$	
x_0	$f(x_0)$		$\delta^2 f(x_0)$		$\delta^4 f(x_0)$
		$\delta f(x_{\frac{1}{2}})$		$\delta^3 f(x_{\frac{1}{2}})$	
x_1	$f(x_1)$		$\delta^2 f(x_1)$		
		$\delta f(x_{\frac{3}{2}})$			
x_2	$f(x_2)$				

§ 5. Funcţii spline

In paragrafele precedente s-au folosit exclusiv polinoamele pentru a aproxima o funcţie dată. Interpolarea polinomială se dovedeşte a fi un proces divergent.

In 1901 Runge a construit exemplul următor: fie $f:[-5,5]\to R$ dată prin

$$f(x) = \frac{1}{1+x^2}$$

şi, pentru un $n\in N^*$, nodurile de interpolare

$$x_i^n = -5 + \frac{10.i}{n}, \qquad i = 0,1,\ldots,n,$$

şi P_n polinomul de interpolare de grad n pe aceste noduri.

Atunci are loc:

$$\|f - P_n\| \xrightarrow[n]{} +\infty ,$$

adică acest proces de interpolare nu converge în norma convergenţei uniforme la funcţia f.

In 1912 Bernstein a considerat funcţia $f(x) = |x|$, $x\in[-1,1]$, nodurile de interpolare

$$x_i^n = -1 + \frac{2i}{n}, \qquad i = 0,1,\ldots,n,$$

(pentru orice $n\in N^*$) şi a arătat că şirul polinoamelor de interpolare corespunzător nu converge uniform la f.

In 1914 Faber a demonstrat următorul rezultat în privinţa divergenţei procesului de interpolare [6]:

Teoremă. Fie, pentru orice $n\in N^*$, nodurile distincte

$$x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}$$

situate în intervalul $[a,b]$. Atunci există o funcţie $f\in C[a,b]$ astfel încît şirul polinoamelor de interpolare corespunzător să nu conveargă uniform la f pe intervalul $[a,b]$.

Aceste fapte au condus la problema găsirii unei alte clase de funcţii de interpolare care să asigure convergenţa procesului de interpolare. In acest sens se introduce clasa aşa-numitelor funcţii spline.

Ideea de bază este următoarea: se împarte intervalul dat în subintervale printr-o reţea de puncte $x_o < x_1 < \dots < x_n$ şi pe fiecare dintre subintervalele rezultate se aproximează funcţia printr-un polinom de interpolare. Se numeşte <u>polinom pe porţiuni</u> de grad $p \in N$ relativ la reţeaua $a = x_o < x_1 < \dots < x_n = b$ o funcţie $\varphi : [a,b] \longrightarrow R$ a cărei restricţie la subintervalele $[x_i, x_{i+1}]$, $i = o,1,\dots,n-1$ este un polinom de grad p. Interpolarea cu polinoame pe porţiuni constă în aproximarea unei funcţii cu un polinom pe porţiuni care să satisfacă anumite condiţii. Cea mai simplă interpolare de acest fel este <u>interpolarea liniară pe porţiuni</u>, care, geometric, constă în a uni prin segmente de drepte punctele succesive din şirul

$$(x_o, f(x_o)),\ (x_1, f(x_1)), \dots, (x_n, f(x_n)) .$$

Aşadar, graficul funcţiei f este aproximat cu linia poligonală cu vîrfurile de mai sus, situate pe graficul lui f. Evident, linia poligonală nu are tangentă în vîrfurile sale. Aceeaşi situaţie apare şi în cazul interpolării cu polinoame pe porţiuni de grad superior. Introducerea funcţiilor spline înlătură acest neajuns.

<u>Definiţie</u>. O <u>funcţie spline de clasă k şi grad p</u> pentru o reţeaua $a = x_o < x_1 < \dots < x_n = b$ este un element $S \in C^k[a,b]$ cu proprietatea că restricţiile $S|[x_i, x_{i+1}]$, $i = o,1,\dots,n-1$, sînt polinoame de grad p.

Foarte utilizate în practică sînt <u>funcţiile spline cubice</u>, care sînt funcţiile spline de clasă 2 şi grad 3. Ca urmare, o funcţie spline cubică S ce interpolează funcţia f pe reţeaua

$a = x_0 < x_1 < \ldots < x_n = b$ satisface condiţiile:

a) $S\big|[x_i, x_{i+1}] = S_i$ (= polinom de gradul 3), $i = 0,1,\ldots,n-1$,

b) $S(x_i) = f(x_i)$, $i = 0,1,\ldots,n$,

c) $S_{i+1}(x_{i+1}) = S_i(x_{i+1})$, $i = 0,1,\ldots,n-2$,

d) $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1})$, $i = 0,1,\ldots,n-2$,

e) $S''_{i+1}(x_{i+1}) = S''_i(x_{i+1})$, $i = 0,1,\ldots,n-2$.

Numărul acestor condiţii este 4n-2. Pe de altă parte, deoarece un polinom de gradul 3 are 4 coeficienţi, funcţia S depinde de 4n parametri. Mai trebuie impuse încă două condiţii. De exemplu, putem considera una din următoarele condiţii la limită:

I $S''(x_0) = S''(x_n) = 0$;

II $S'(x_0) = f'(x_0)$ şi $S'(x_n) = f'(x_n)$.

O funcţie spline cubică ce satisface condiţia I se numeşte <u>funcţie spline naturală</u> cu nodurile $x_0, x_1, \ldots, x_n$. Aproximarea unei funcţii cu o funcţie spline cubică ce satisface condiţia II este, în general, mai bună decît aproximarea ei cu o funcţie spline naturală; aceasta datorită faptului că în condiţia II sînt incluse informaţii suplimentare asupra funcţiei date. Pe de altă parte, este necesar să avem în acest caz valorile derivatei funcţiei f în extremităţile intervalului sau aproximaţii cît mai bune ale lor.

Ne propunem să construim o funcţie spline naturală. Construcţia unei funcţii spline cubice ce satisface condiţia II se face la fel (şi coincide cu cea de mai jos pînă în momentul utilizării condiţiei I).

<u>Teorema 5.1. Dacă se cunosc valorile unei funcţii $f:[a,b] \to R$ în punctele reţelei $a = x_0 < x_1 < \ldots < x_n = b$, atunci există o funcţie spline naturală pentru această reţea, unică, ce o interpolează.</u>

Demonstrație. Să presupunem că S este o funcție spline cubică pentru rețeaua dată ce interpolează funcția f, adică funcția $S : [a,b] \to R$ satisface condițiile a) - e). Să scriem polinomul S_i sub forma:

$$(S) \quad S_i(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3, \quad i=0,1,\ldots,n-1,$$

și să vedem ce condiții revin coeficienților a_i, b_i, c_i, d_i, $i=0,1,\ldots,n-1$, din impunerea condițiilor a) - e). Din a) și b) și avînd în vedere (S) rezultă că

$$a_i = S_i(x_i) = S(x_i) = f(x_i), \quad i = 0,1,\ldots,n-1,$$

adică pentru coeficienții a_i avem:

$$(5.1) \qquad a_i = f(x_i), \qquad i = 0,1,\ldots,n-1 .$$

Acum, ținînd seama de c), obținem:

$$a_{i+1} = S_{i+1}(x_{i+1}) = S_i(x_{i+1}) = a_i + b_i(x_{i+1}-x_i) + c_i(x_{i+1}-x_i)^2 + \\ + d_i(x_{i+1}-x_i)^3, \quad i = 0,1,\ldots,n-2,$$

sau, notînd $h_i = x_{i+1} - x_i$ $(i = 0,1\ldots,n-1)$,

$$(5.2) \qquad a_{i+1} = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \quad i = 0,1,\ldots,n-2 .$$

Dacă definim

$$(5.3) \qquad a_n = f(x_n),$$

vom avea și relația

$$a_n = f(x_n) = S(x_n) = S_{n-1}(x_n) = a_{n-1} + b_{n-1}h_{n-1} + c_{n-1}h_{n-1}^2 + d_{n-1}h_{n-1}^3,$$

deci a_n este de forma (5.2). Ca urmare, avem

$$(5.4) \qquad a_{i+1} = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \quad i = 0,1,\ldots,n-1 .$$

Derivata polinomului S_i este

$$(S') \qquad S_i'(x) = b_i + 2c_i(x-x_i) + 3d_i(x-x_i)^2, \quad i = 0,1,\ldots,n-1 .$$

Atunci $b_i = S_i'(x_i)$, $i = 0,1,\ldots,n-1$, și apelînd la condiția

d) vom obține:

(5.5) $$b_{i+1} = S'_{i+1}(x_{i+1}) = S'_i(x_{i+1}) = b_i + 2c_ih_i + 3d_ih_i^2, \quad i=0,1,\ldots,n-2.$$

Dacă definim

$$b_n = S'(x_n),$$

vom avea și relația

$$b_n = S'(x_n) = S'_{n-1}(x_n) = b_{n-1} + 2c_{n-1}h_{n-1} + 3d_{n-1}h_{n-1}^2$$

care este de forma (5.5) și, deci, are loc

(5.6) $$b_{i+1} = b_i + 2c_ih_i + 3d_ih_i^2, \quad i = 0,1,\ldots,n-1.$$

In sfîrșit, derivata secundă a polinomului S_i este

(S") $$S''_i(x) = 2c_i + 6d_i(x-x_i), \quad i = 0,1,\ldots,n-1.$$

Atunci $c_i = \frac{1}{2} S''_i(x_i)$, $i = 0,1,\ldots,n-1$, și apelînd la e), obținem:

(5.7) $$c_{i+1} = \frac{1}{2} S''_{i+1}(x_{i+1}) = \frac{1}{2} S''_i(x_{i+1}) = c_i + 3d_ih_i, \quad i=0,1,\ldots,n-2.$$

Dacă definim

(5.8) $$c_n = \frac{1}{2} S''(x_n),$$

vom obține

$$c_n = \frac{1}{2} S''(x_n) = \frac{1}{2} S''_{n-1}(x_n) = c_{n-1} + 3d_{n-1}h_{n-1},$$

care este de forma (5.7) și, ca urmare, avem

(5.9) $$c_{i+1} = c_i + 3d_ih_i, \quad i = 0,1,\ldots,n-1.$$

Din (5.9), (5.4) și (5.6) obținem:

(5.10) $$d_i = \frac{1}{3h_i}(c_{i+1} - c_i), \quad i = 0,1,\ldots,n-1;$$

(5.11) $$a_{i+1} = a_i + b_i h_i + \frac{h_i^2}{3}(2c_i + c_{i+1}), \quad i=o,1,\ldots,n-1;$$

(5.12) $$b_{i+1} = b_i + h_i(c_i + c_{i+1}), \quad i = o,1,\ldots,n-1.$$

Din (5.11) rezultă că:

(5.13) $$b_i = \frac{1}{h_i}(a_{i+1} - a_i) - \frac{h_i}{3}(2c_i + c_{i+1}), \quad i = o,1,\ldots,n-1,$$

iar (5.12) se mai poate scrie

$$b_i = b_{i-1} + h_{i-1}(c_{i-1} + c_i), \quad i = 1,2,\ldots,n .$$

Inlocuind aici b_i şi b_{i-1} daţi de (5.13), sîntem conduşi la

$$\frac{1}{h_i}(a_{i+1} - a_i) - \frac{h_i}{3}(2c_i + c_{i+1}) = \frac{1}{h_{i-1}}(a_i - a_{i-1}) - \frac{h_{i-1}}{3}(2c_{i-1} + c_i) + h_{i-1}(c_{i-1} + c_i)$$

sau

(5.14) $$h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_i c_{i+1} = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}), \quad i=1,2,\ldots,n-1.$$

Condiţia I conduce la

(5.15) $$c_o = c_n = o$$

(într-adevăr, din (5,8) avem $c_n = \frac{1}{2}S''(x_n) = o$ şi, pe de altă parte, $o = S''(x_o) = S_o''(x_o) = 2c_o + 6d_o(x_o - x_o) = 2c_o$, adică $c_o = o$). Deci numerele c_i, $i = o,1,\ldots,n$, sînt, în mod necesar soluţii ale sistemului de n+1 ecuaţii cu n+1 necunoscute (5.14) + (5.15), unde a_i, $i = o,1,\ldots,n$ sînt daţi de (5.1) şi (5.3). Dacă acest sistem ar avea o soluţie unică, atunci relaţiile (5.1), , (5.13) şi (5.10) ar defini numerele a_i, b_i, d_i, $i = o,1,\ldots,n-1$ şi prin intermediul formulei (S) am fi conduşi la n polinoame S_i, $i = o,1,\ldots,n-1$. In acest caz, funcţia S definită prin $S|[x_i, x_{i+1}] = S_i$ ar fi unica funcţie spline cubică ce satisface condiţiile a) - e) şi I şi demonstraţia teoremei ar fi încheiată.

Să arătăm, aşadar, că sistemul format din ecuaţiile (5.14) şi (5.15) are soluţie unică. Vom face mai mult, şi anume, vom determina această soluţie ceea ce va permite degajarea algoritmului de calcul al funcţiei spline naturale căutate. Să notăm

$$(5.16)\qquad \alpha_i = \frac{3}{h_i}(a_{i+1} - a_i) - \frac{3}{h_{i-1}}(a_i - a_{i-1}) =$$

$$= \frac{3}{h_i h_{i-1}}\left[h_i a_{i-1} - (h_{i-1} + h_i)a_i + h_{i-1}a_{i+1}\right],\ i=o,1,\dots,n-1.$$

Atunci sistemul (5.14) + (5.15) se scrie:

$$(5.17)\quad \begin{cases} E_o: & c_o = o \\ E_1: & h_o c_o + 2(h_o+h_1)c_1 + h_1 c_2 = \alpha_1 \\ E_2: & h_1 c_1 + 2(h_1+h_2)c_2 + h_2 c_3 = \alpha_2 \\ \text{-----} & \text{-----} \\ E_i: & h_{i-1}c_{i-1} + 2(h_{i-1}+h_i)c_i + h_i c_{i+1} = \alpha_i \\ E_{i+1}: & h_i c_i + 2(h_i+h_{i+1})c_{i+1} + \\ & + h_{i+1}c_{i+2} = \alpha_{i+1} \\ \text{-----} & \text{-----} \\ E_{n-1}: & h_{n-2}c_{n-2} + 2(h_{n-2}+h_{n-1})c_{n-1} + \\ & + h_{n-1}c_n = \alpha_{n-1} \\ E_n: & c_n = o \end{cases}$$

Matricea sistemului (5.17) este tridiagonală şi este inversabilă, deoarece diagonala domină strict asupra liniilor. Ca urmare (5.17) are soluţie unică.

Să observăm că putem elimina c_o din ecuaţia E_1 dacă ţinem seama de E_o; notăm E_1' ecuaţia ce rezultă. Printr-o combinaţie elementară a ecuaţiilor E_1' şi E_2 se poate elimina c_1 din E_2; notăm E_2' ecuaţia obţinută. Procedînd în acest mod, prin combinarea ecuaţiilor E_i' (ce nu conţine c_{i-1}) şi E_{i+1} se elimină c_i din E_{i+1}. Ecuaţia obţinută, notată E_{i+1}' conţine necunoscutele c_{i+1} şi c_{i+2}, necunoscuta c_{i+2} avînd coeficientul h_{i+1} ca şi în vechea ecuaţie E_{i+1}. Ultima ecuaţia E_n rămîne neschimbată. In

felul acesta, sistemul (5.17) este adus la forma

$$(5.18)\qquad \begin{cases} E_o : & c_o = o \\ E_i' : & \lambda_i c_i + h_i c_{i+1} = \beta_i, \quad i = 1,2,\dots,n-1, \\ E_n : & c_n = o \end{cases}$$

unde

$$(5.19)\qquad \begin{cases} \lambda_1 = 2(h_o + h_1) \\ \beta_1 = \alpha_1 \end{cases}$$

iar λ_i, β_i $(i = 1,2,\dots,n-1)$ se determină prin recurență:

$$(5.20)\qquad \begin{cases} \lambda_{i+1} = 2(h_i + h_{i+1}) - \dfrac{h_i^2}{\lambda_i} \\ \beta_{i+1} = \alpha_{i+1} - \dfrac{h_i}{\lambda_i}\beta_i \end{cases}, \quad i=1,2,\dots,n-2 .$$

Intr-adevăr, din cele spuse mai sus, $E_{i+1}' = E_{i+1} - \dfrac{h_i}{\lambda_i} E_i'$, $i = 1,2,\dots,n-1$, de unde rezultă relațiile (5.20). Faptul că $\lambda_i \neq o$, $i = 1,2,\dots,n-1$, decurge din inegalitățile $\lambda_i > h_i$, $i = 1,2,\dots,n-1$, a căror verificare este imediată. Soluția sistemului (5.18) este dată de

$$(5.21)\qquad \begin{cases} c_n = o \\ c_i = \dfrac{1}{\lambda_i}(\beta_i - h_i c_{i+1}), \quad i=n-1,n-2,\dots,2,1, \\ c_o = o , \end{cases}$$

unde λ_i și β_i sînt determinați prin (5.19) și (5.20).

<u>Algoritmul interpolării cu funcții spline naturale.</u> Fie f o funcție cu valori date în nodurile $x_o < x_1 < \dots < x_n$. Funcția spline naturală pentru această rețea se construiește astfel:

<u>Pasul 1</u>. Calculează, pentru $i = 1,2,\dots,n-1$,

$$\alpha_i = \frac{3\left[f(x_{i-1})(x_{i+1}-x_i) - f(x_i)(x_{i+1}-x_{i-1}) + f(x_{i+1})(x_i - x_{i-1})\right]}{(x_{i+1}-x_i)(x_i - x_{i-1})} .$$

<u>Pasul 2</u>. Calculează $\lambda_1 = 2(x_2 - x_o)$, $\beta_1 = \alpha_1$.

Pasul 3. Pune $i = 1$

Pasul 4. Calculează

$$\lambda_{i+1} = 2(x_{i+2}-x_i) - \frac{(x_{i+1}-x_i)^2}{\lambda_i},$$

$$\beta_{i+1} = \alpha_{i+1} - \frac{(x_{i+1}-x_i)\beta_i}{\lambda_i}$$

Pasul 5. Adună 1 la i.

Pasul 6. Dacă $i < n-1$, treci la Pasul 4. Dacă $i = n-1$, treci la Pasul 7.

Pasul 7. Pune $c_n = 0$.

Pasul 8. Pune $j = n-1$.

Pasul 9. Calculează

$$c_j = \frac{\beta_j - (x_{j+1}-x_j)c_{j+1}}{\lambda_j},$$

$$b_j = \frac{f(x_{j+1}) - f(x_j)}{x_{j+1}-x_j} - \frac{(x_{j+1}-x_j)(2c_j + c_{j+1})}{3},$$

$$d_j = \frac{c_{j+1}-c_j}{3(x_{j+1}-x_j)}.$$

Pasul 10. Scade 1 din j.

Pasul 11. Dacă $j > 0$, treci la Pasul 9. Dacă $j = 0$, pune

$$c_0 = 0, \quad b_0 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} - \frac{(x_1-x_0)c_1}{3}, \quad d_0 = \frac{c_1}{3(x_1-x_0)}.$$

Pasul 12. Procedura este completă. Funcţia spline naturală este dată pe $[x_i, x_{i+1}]$ de

$$S_i(x) = f(x_i) + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3.$$

§ 6. Exerciţii

1. Să se arate că

$$f[x_0,x_1,\dots,x_n]=\frac{V(x_0,x_1,\dots,x_n;\ f)}{V(x_0,x_1,\dots,x_n)}$$

unde $V(x_0,x_1,\dots,x_n)$ este determinantul Vandermonde al numerelor $x_0,x_1,\dots,x_n$ şi $V(x_0,x_1,\dots,x_n;\ f)$ este un determinant de ordin n+1 format cu aceleaşi linii ca şi determinantul Vandermonde, cu excepţia ultimei linii care se înlocuieşte cu $(f(x_0),\ f(x_1),\dots$ $\dots,\ f(x_n))$.

2. Fie funcţiile $f:I\longrightarrow R$, $g:I\longrightarrow R$ şi nodurile $x_0,x_1,\dots$ $\dots,x_n$ aparţinînd intervalului I. Are loc egalitatea:

$$(f.g)[x_0,x_1,\dots,x_n]=\sum_{i=0}^{n} f[x_0,x_1,\dots,x_i]\cdot g[x_i,x_{i+1},\dots,x_n]\ .$$

3. Fie funcţia $f:I\longrightarrow R$, nodurile $x_0,x_1,\dots,x_n$ din I şi numărul natural $i<n$. Dacă $g:I\longrightarrow R$ este definită prin

$$g(x)=f[x_{i+1},\dots,x_n,x]\ ,$$

atunci are loc formula:

$$f[x_0,x_1,\dots,x_n]=g[x_0,x_1,\dots,x_i]\ .$$

4. Să se arate că:

a) $\Delta-\nabla=\Delta\nabla$;

b) $\Delta\sqrt{f_k}=\dfrac{\Delta f_k}{\sqrt{f_k}+\sqrt{f_{k+1}}}$;

c) $\sum_{k=0}^{n-1}\Delta^2 f_k=\Delta f_n-\Delta f_0$.

5. Să se arate că:

$$\sum_{i=0}^{n} L_i(x)=1$$

pentru orice x, unde

$$L_i(x) = \frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)} .$$

6. Fie funcţia $f(x) = \sqrt[3]{x}$ şi nodurile $x_0 = 27$, $x_1 = 64$ şi $x_2 = 125$.

a) Să se scrie formula de interpolare a lui Lagrange pe aceste noduri.

b) Să se aproximeze numărul $\sqrt[3]{100}$ cu ajutorul acestei formule şi să se evalueze eroarea comisă.

Aceleaşi chestiuni relativ la $f(x) = 3^x$, nodurile $x_0 = -2$, $x_1 = -1$, $x_2 = 0$, $x_3 = 1$, $x_4 = 2$ şi punctul $x = \sqrt{3}$.

7. Funcţia $y = f(x)$ este dată în punctele (7,3), (8,1), (9,1) şi (10,9). Să se găsească valoarea lui y pentru x = 9,5 utilizînd:

a) formula de interpolare a lui Lagrange;

b) formula de interpolare a lui Newton.

8. Tabelul de date

x	1,3	1,6	1,9
f(x)	0,6200860	0,4554022	0,2818186
f'(x)	-0,5220232	-0,5698959	-0,5811571

corespunde funcţiei Bessel $J_1(x)$. Să se determine o aproximaţie pentru $J_1(1,5)$.

9. Utilizînd formula de interpolare a lui Hermite să se aproximeze f(2,5), dat fiind tabelul

x	2,2	2,4	2,6
f(x)	0,5207843	0,5104147	0,4813306
f'(x)	-0,0014878	-0,1004889	-0,1883635

10. Fie dat tabelul de valori

x	0,0	0,2	0,4	0,6	0,8
f(x)	1,00000	1,22140	1,49182	1,82212	2,22554

Să se calculeze, alegînd formula de interpolare convenabilă, următoarele valori: a) f(0,05), b) f(0,43) şi c) f(0,75).

11. Fie dat tabelul de valori

x	0,0	0,1	0,3	0,6	1,0
f(x)	-7,00000	-5,89483	-5,65014	-5,17788	-4,28172

a) Să se alcătuiască schema cu diferenţe divizate corespunzătoare tabelului.

b) Să se scrie formula de interpolare a lui Newton pe aceste noduri.

12. Funcţia f este dată de tabelul de valori

x	0,01	0,02	0,03	0,04	0,05
f(x)	98,4342	48,4392	31,7775	23,4492	18,4542

Să se găsească f(0,0341).

13. Funcţia

$$f(x) = \int_x^{\infty} \frac{e^{-t}}{t}\, dt$$

este dată de

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06
f(x)	∞	4,0379	3,3547	2,9591	2,6813	2,4679	2,2953

Să se aproximeze f(0,0378).

14. Se consideră nodurile $a = x_0 < x_1 < x_2 < x_3 = b$ şi funcţia

$$S(x) = c_1(x-x_1)_+^3 + c_2(x-x_2)_+^3$$

cu convenţia ca

$$(x-t)_+^3 = \begin{cases} (x-t)^3 & , x > t , \\ 0 & , x \le t . \end{cases}$$

Să se arate că $S \in C^2[a,b]$.

15. Se consideră polinomul $P(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ şi nodurile $a = x_0 < x_1 < \ldots < x_n = b$. Să se arate că funcţia

$$S(x) = P(x) + \sum_{k=1}^{n-1} c_k (x - x_k)_+^3$$

este o funcţie spline cubică.

16. Fie dat tabelul de valori

x	o	o,1	o,2	o,3	o,4	o,5	o,6	o,7	o,8	o,9	1
$f(x)$	o	-o,1	-o,3	-o,5	-o,4	-o,2	o	o,2	o,3	o,7	o,9

Să se găsească funcţia spline naturală care interpolează f pe nodurile $x_i = $ o,i, $i = $ o,1,2,...,1o.

Capitolul VI

INTEGRARE NUMERICA

§ 1. Generalităţi asupra formulelor de cuadratură

Problema abordată în acest capitol este aceea a aproximării integralelor definite.

Fie f o funcţie integrabilă pe intervalul $I \subset R$, $[a,b] \subset I$ şi

(1.1) $$\int_a^b p(x)\, f(x)\, dx \,,$$

unde $p : [a,b] \longrightarrow R$ este funcţia pondere, o funcţie fixată pe care o presupunem pozitivă şi continuă. Foarte adesea apare situaţia cînd nu se poate calcula integrala (1.1) utilizînd formula lui Newton-Leibniz sau, dacă am cunoaşte o primitivă a funcţiei de integrat, este anevoios calculul variaţiei acesteia. De asemenea, adesea se cunosc valorile funcţiei numai într-un număr finit de puncte ale intervalului dat, adică funcţia f este tabelată. In aceste situaţii se utilizează metode numerice pentru determinarea unei valori aproximative a integralei (1.1).

Problema integrării numerice a unei funcţii f constă în aflarea unei valori aproximative a integralei (1.1) cu ajutorul unui număr finit de valori ale acestei funcţii.

O formulă de integrare numerică sau de cuadratură este o egalitate de forma :

(1.2) $$\int_a^b p(x)f(x)dx = A_0f(x_0)+ A_1f(x_1)+\ldots+A_nf(x_n)+R_n(f),$$

unde punctele $x_o, x_1, \ldots, x_n \in I$ şi numerele $A_o, A_1, \ldots, A_n \in R$ se numesc respectiv nodurile şi coeficienţii formulei de cuadratură iar $E_n(f)$ (sau $E_n(f,[a,b])$, atunci cînd este necesară indicarea intervalului de integrare) este restul formulei de cuadratură. Ca urmare, o valoare aproximativă a integralei (1.1) este dată de

$$\int_a^b p(x)f(x)dx \approx A_of(x_o)+ A_1f(x_1)+\ldots+ A_nf(x_n) \tag{1.3}$$

eroarea acestei egalităţi aproximative fiind tocmai $E_n(f)$. Aşadar, restul $E_n(f)$ reprezintă eroarea de trunchiere a formulei de cuadratură (1.2).

Să observăm că suma din membrul drept din (1.2) sau (1.3) comportă 2n+2 parametri: n+1 noduri şi n+1 coeficienţi. Diferite tipuri de formule de cuadratură corespund la moduri diferite de alegere a nodurilor x_i, $i=\overline{o,n}$, coeficienţilor A_i, $i=\overline{o,n}$, cît şi ponderei p. Un mijloc de obţinere a unei formule de cuadratură este următorul: se aproximează funcţia f cu o funcţie g şi se consideră egalitatea aproximativă

$$\int_a^b p(x)f(x)dx \approx \int_a^b p(x)g(x)dx$$

(bineînţeles, se contează pe faptul că integrala din membrul al doilea este uşor de calculat). In particular, şi cel mai adesea, se ia drept funcţie g o funcţie de interpolare şi chiar un polinom de interpolare. In această ordine de idei, în cele ce urmează vom avea în vedere numai formule de cuadratură interpolative, adică acelea ce se obţin prin integrarea unei formule de interpolare relativ la funcţia f (după ce această din urmă formulă a fost înmulţită cu ponderea p).

Se numeşte grad de exactitate al formulei de cuadratură (1.2) numărul natural m ce satisface la condiţiile: 1°. $E_n(P_k)=o$ pentru orice polinom P_k de grad k mai mic sau cel mult egal cu m;

2° există un polinom de grad m+1, fie el P_{m+1}, astfel încît $E_n(P_{m+1}) \neq o$. Dacă convenim să spunem că (1.2) este exactă pentru funcţia φ dacă $E_n(\varphi) = o$, putem reformula definiţia dată în felul următor: m este gradul de exactitate al formulei (1.2) dacă această formulă este exactă pentru orice polinom de grad cel mult m şi m este cel mai mare număr natural cu această proprietate. Mai mult, este uşor de constatat că m este gradul de exactitate al formulei de cuadratură (1.2) dacă şi numai dacă

(1.4) $\quad E_n(x^k) = o \quad (k = o,1,\dots,m)$ şi $E_n(x^{m+1}) \neq o$.

Intr-adevăr, din liniaritatea în raport cu f atît a integralei cît şi a sumei $A_of(x_o)+\dots+A_nf(x_n)$ ce figurează în egalitatea (1.2) rezultă că $E_n(f)$ este la rîndul ei liniară şi deci afirmaţia de mai sus este corectă.

<u>Exerciţii</u>. Fie funcţia $f: [-1,2] \longrightarrow R$ şi formula de cuadratură

$$\int_{-1}^{+1} f(x)dx = A_of(-1) + A_1f(o) + A_2f(2) + E_2(f) .$$

1) Să se determine coeficienţii A_o, A_1, A_2 astfel ca această formulă să fie exactă pentru toate polinoamele de grad mai mic sau cel mult egal cu 2.

2) Să se determine gradul de exactitate al formulei.

<u>Soluţie</u>. 1) Impunînd condiţiile $E_2(1)= o$, $E_2(x)= o$ şi $E_2(x^2)= o$, se obţine sistemul liniar

$$\begin{cases} A_o + A_1 + A_2 = 2, \\ -A_o \qquad +2A_2 = o , \\ A_o \qquad +4A_2 = \frac{2}{3} , \end{cases}$$

a cărui soluţie este $A_o = \frac{2}{9}$, $A_1 = \frac{5}{3}$ şi $A_2 = \frac{1}{9}$. Formula de cuadratură devine

(1.5) $$\int_{-1}^{+1} f(x)dx = \frac{2}{9} f(-1) + \frac{5}{3} f(o) + \frac{1}{9} f(2) + E_2(f).$$

2) Intrucît

$$E_2(x^3) = \int_{-1}^{+1} x^3 dx - \left[\frac{2}{9}(-1)^3 + \frac{5}{3}.o + \frac{1}{9}.2^3\right] = \frac{2}{3} \neq o \quad ,$$

rezultă că formula de cuadratură (1.5) este exactă de gradul 2.

Observaţie. Nodurile unei formule de cuadratură pot să nu aparţină intervalului de integrare.

O limitare a gradului de exactitate este dată de

Teorema 1.1. Gradul de exactitate m al unei formule de cuadratură interpolative pe n+1 noduri satisface condiţia

(1.6) $$n \leq m \leq 2n+1 \quad .$$

Demonstraţie. Prin ipoteză formula de cuadratură derivă dintr-o formulă de interpolare pe n+1 noduri. Stim că o formulă ca aceasta din urmă este exactă pentru orice polinom de grad mai mic sau cel mult egal cu n. Se vede imediat că această proprietate se transmite şi formulei de cuadratură interpolative. Ca urmare $m \geqslant n$.

Pe de altă parte, este uşor de indicat un polinom P(x) de grad 2n+2 astfel încît $E_n(P) \neq o$. Intr-adevăr, fie

$$P(x) = (x-x_o)^2(x-x_1)^2. \;\ldots\; .(x-x_n)^2$$

(amintim că $x_o, x_1, \ldots, x_n$ sînt tocmai cele n+1 noduri). Se verifică uşor că

$$\int_a^b p(x).P(x)dx > o \quad \text{şi} \quad \sum_{i=o}^{n} A_i P(x_i) = o$$

oricare ar fi coeficienţii A_i, $i = o,1,\ldots,n$. Aşadar, $E_n(P) \neq o$ şi $n \leq 2n+1$, q.e.d.

Observaţie. Se pune problema dacă există formule de cuadratură de tip interpolativ al căror grad de exactitate să fie minim (adică egal cu n) sau maxim (adică egal cu 2n+1). Răspun-

sul este afirmativ, după cum se va vedea în paragrafele următoare.

§ 2. Formule de cuadratură Newton-Côtes simple

Aceste formule de cuadratură sînt interpolative - se obţin prin integrarea formulei de interpolare a lui Lagrange. Ele au particularitatea că nodurile de cuadratură sînt echidistante. Gradul lor de exactitate este n, în general (subînţelegem că numărul nodurilor considerate este n+1), adică se atinge extrema stîngă în formula (1.6).

Fie funcţia $f:I \longrightarrow R$, intervalul $[a,b] \subset I$ şi reţeaua de noduri echidistante

(2.1) $$x_i = x_o + ih \in I \ , \quad i = o,1,\dots,n,$$

numărul h fiind pasul reţelei. Formulele de cuadratură Newton - Côtes se clasifică în:

A. formule închise - în acest caz $x_o = a$ şi $x_n = b$ şi deci $h = \frac{b-a}{n}$;

B. formule deschise - în acest caz $x_o = a+h$ şi $x_n = b-h$ şi deci $h = \frac{b-a}{n+2}$.

Odată fixate nodurile x_i, $i = o,1,\dots,n$, mai rămîne de determinat, în ambele aceste cazuri, cei n+1 coeficienţi $A_o, A_1, \dots, A_n$ ai formulei de cuadratură (1.2).

A. Formula de cuadratură Newton-Côtes închise

Să scriem formula de interpolare a lui Lagrange pe nodurile x_i, $i = o,1,\dots,n$ date de (2.1); avem:

(2.2) $$f(x) = L(x_o,x_1,\dots,x_n;\ f|x) + R_n(x) \ ,$$

unde

(2.3) $$L(x_o,x_1,\dots,x_n;\ f|x) = \sum_{i=o}^{n} l_i(x)\, f(x_i)$$

cu

$$(2.4)\qquad l_i(x)=\frac{(x-x_0)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}\ ,$$

$$i=o,1,\dots,n\ ,$$

şi

$$(2.5)\qquad R_n(x)=(x-x_0)\dots(x-x_n)\ f[x,x_0,\dots,x_n]$$

sau, în ipoteza că $f\in C^{n+1}[a,b]$,

$$(2.6)\qquad R_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)\dots(x-x_n),\quad \xi=\xi(x)\in(a,b)$$

Inmulţind (2.2) cu ponderea p şi integrînd pe intervalul $[a,b]$, se obţine formula de cuadratură

$$(2.7)\qquad \int_a^b p(x)f(x)dx=\sum_{i=o}^{n}A_if(x_i)+E_n(f)\ ,$$

unde, ţinînd seama de (2.3),

$$(2.8)\qquad A_i=\int_a^b p(x)\ l_i(x)\ dx\ ,\qquad i=o,1,\dots,n\ ,$$

şi, în virtutea relaţiei (2.5),

$$(2.9)\qquad E_n(f)=\int_a^b p(x)(x-x_0)\dots(x-x_n)\ f[x,x_0,\dots,x_n]dx$$

sau, graţie ipotezei suplimentare de mai sus şi ţinînd seama de (2.6),

$$(2.10)\qquad E_n(f)=\frac{1}{(n+1)!}\int_a^b p(x)(x-x_0)\dots(x-x_n)\ f^{(n+1)}(\xi)\ dx$$

Se constată, urmărind formulele (2,8), că valorile coeficienţilor A_i nu depind de funcţia f.

Observaţie. Intrucît diferenţa divizată ce apare în (2.9) se anulează pentru orice polinom de grad $\leq n$, rezultă că restul E_n este nul pentru astfel de polinoame şi deci gradul de exactitate al formulei de cuadratură (2.7) este cel puţin n.

In continuare vom considera cazul particular cînd ponderea $p(x)=1$, $x\in[a,b]$. In acest cadru se pot da expresii simple pentru coeficienţii şi restul formulei de cuadratură (2.7). Să

efectuăm schimbarea de variabilă

(2.11) $x = x_o + th$, $t \in [o,n]$,

în integrala din (2.8). Ţinînd seama de expresia (2.4) pentru polinomul $L_i(x)$ şi observînd că $x-x_k = (t-k)h$ şi $x_j - x_k = (j-k)h$ $(j,k = o,1,\dots,n)$, se obţine :

$$A_i = \frac{(-1)^{n-i}h}{i!(n-i)!}\int_o^n \frac{t(t-1)\dots(t-n)}{t-i}\,dt , \quad i = o,1,\dots,n ,$$

sau

(2.12) $A_i = (b-a)\,H_i$, $i = o,1,\dots,n$,

cu

(2.13) $$H_i = \frac{1}{n}\,\frac{(-1)^{n-i}}{i!(n-i)!}\int_o^n \frac{t(t-1)\dots(t-n)}{t-i}\,dt .$$

Constantele H_i, $i = o,1,\dots,n$, se numesc <u>coeficienţii lui Côtes</u>. Ca urmare, formula de cuadratură (2.7) ia forma:

(2.14) $$\int_a^b f(x)dx = (b-a)\sum_{i=o}^{n} H_i f(x_i) + E_n(f) ,$$

pentru restul $E_n(f)$ avînd deocamdată expresiile (2.9) sau, eventual, (2.10) în care luăm $p(x) = 1$.

Formulele (2.13) permit să remarcăm că aceşti coeficienţi H_i nu depind nici de funcţia f, nici de intervalul $[a,b]$, ci numai de n, adică de numărul nodurilor. De asemenea, coeficienţii lui Côtes au următoarele proprietăţi:

(i) $\sum_{i=o}^{n} H_i = 1$,

(ii) $H_i = H_{n-i}$, $i = o,1,\dots,n$

Intr-adevăr, dacă în (2.14) considerăm funcţia $f(x) = 1$, obţinem (i) şi dacă în (2.13) efectuăm schimbarea de variabilă $n-t = u$, obţinem (ii). Aceste proprietăţi cît şi formulele (2.13) permit calcularea cu uşurinţă a coeficienţilor lui Côtes (pentru

diversele valori ale lui n) şi înregistrarea lor într-un tabel. Lectorul interesat poate verifica singur corectitudinea valorilor acestor coeficienţi trecute în tabelul de mai jos, pentru n pînă la 5.

Tabel

Coeficienţii lui Côtes - formule N-C închise

h \ H	H_0	H_1	H_2	H_3	H_4	H_5	H_6
1	$\frac{1}{2}$	$\frac{1}{2}$					
2	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$				
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$			
4	$\frac{7}{90}$	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{7}{90}$		
5	$\frac{19}{288}$	$\frac{75}{288}$	$\frac{50}{288}$	$\frac{50}{288}$	$\frac{75}{288}$	$\frac{19}{288}$	

Să revenim la restul $E_n(f)$ al formulei de cuadratură (2.14). Să impunem funcţiei f ipoteza suplimentară $f \in C^{n+1}[a,b]$. Atunci, conform cu (2.10), avem:

$$(2.15) \qquad E_n(f) = \frac{1}{(n+1)!}\int_a^b f^{(n+1)}(\xi)(x-x_0)\dots(x-x_n)dx$$

(ξ depinzînd de x). De aici rezultă imediat o evaluare a erorii de trunchiere a formulei (2.14):

$$(2.16) \qquad |E_n(f)| \le \frac{1}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)| \cdot \int_a^b |(x-x_0)\dots(x-x_n)| dx .$$

Se pot aduce îmbunătăţiri importante formei (2.15) pentru restul $E_n(f)$. Acest fapt este cuprins în teorema următoare, care dă precizări definitive în privinţa formulelor de cuadratură Newton-Côtes de tip închis.

<u>Teorema 2.1. Fie $f : [a,b] \to R$, $x_0 = a$, $x_n = b$, $x_i = x_0 + ih$, $(i = 0,1,\dots,n)$ şi $h = \frac{b-a}{n}$. Atunci</u> ,

(i) dacă n este par şi $f \in C^{n+2}[a,b]$, există un punct $\xi \in (a,b)$ astfel încît

$$(2.17)\quad \int_a^b f(x)dx = (b-a)\sum_{i=0}^{n} H_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!}\int_0^n t^2(t-1)\dots(t-n)dt\ ;$$

(ii) dacă n este impar şi $f \in C^{n+1}[a,b]$, există un punct $\xi \in (a,b)$ astfel încît

$$(2.18)\quad \int_a^b f(x)dx = (b-a)\sum_{i=0}^{n} H_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!}\int_0^n t(t-1)\dots(t-n)dt\ ,$$

unde H_i $(i=0,1,\dots,n)$ sînt coeficienţii lui Côtes daţi de (2.13).

Lectorul iubitor de rigoare poate găsi în paragraful următor demonstraţia detaliată a formelor pentru rest date de Teorema 2.1. Acest paragraf se poate însă omite de către cei grăbiţi, fără consecinţă asupra consideraţiilor următoare.

Observaţie. In conformitate cu teorema precedentă, dacă n este par, gradul de exactitate al formulei de cuadratură este n+1 (deşi polinomul de interpolare utilizat avea gradul $\leq n$); dacă însă n este impar, gradul de exactitate este egal cu n. Ca urmare, dacă n este par, gradul de exactitate nu se măreşte prin adăugarea unui singur nod, acest fapt va avea loc numai atunci cînd adăugăm cel puţin două noduri noi.

Prezentăm cîteva formule Newton-Côtes simple şi de tip închis pentru valori mici ale lui n (se obţin din (2.17) sau (2.18) prin calcule imediate).

n = 1 - formula trapezului

$$(2.19)\quad \int_a^b f(x)dx = \frac{h}{2}\left[f(x_0)+f(x_1)\right] - \frac{h^3}{3} f''(\xi)$$

unde $x_0 = a$, $x_1 = b$, $h = b-a$ şi $x_0 < \xi < x_1$. Semnificaţia geometrică a părţilor acestei formule este indicată în fig.2.1.

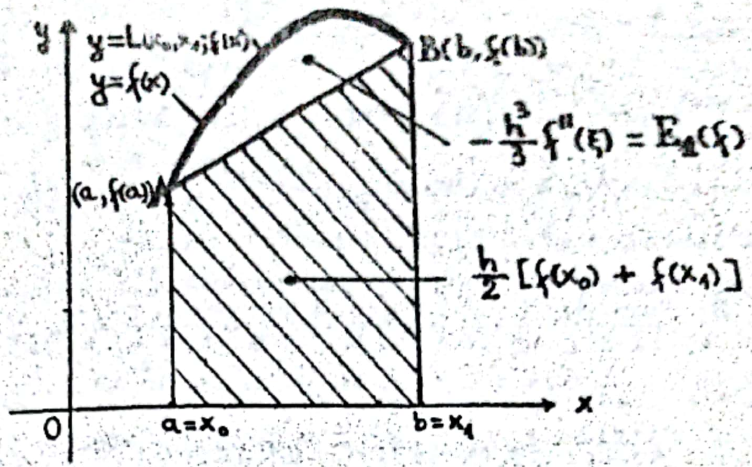


Fig.2.1.

Integrala din membrul întîi se aproximează cu aria trapezului haşurat; aria suprafeţei dintre coarda AB şi arcul AB al graficului funcţiei $y = f(x)$ este restul formulei.

<u>n = 2</u> - <u>formula lui Simpson (formula parabolei)</u>

(2.20) $$\int_a^b f(x)dx = \frac{h}{3}\left[f(x_0)+ 4f(x_1)+ f(x_2)\right]- \frac{h^5}{90} f^{(4)}(\xi) ,$$

unde $x_0= a$, $x_2= b$, $h = (b-a)/2$ şi $x_0 < \xi < x_2$. Integrala este aproximată de aria suprafeţei plane cuprinsă între axa Ox, dreptele $x = a$ şi $x = b$ şi graficul parabolei cu axa paralelă cu Oy şi care trece prin punctele $(x_0,f(x_0))$, $(x_1,f(x_1))$ şi $(x_2,f(x_2))$ (fig.2.2).

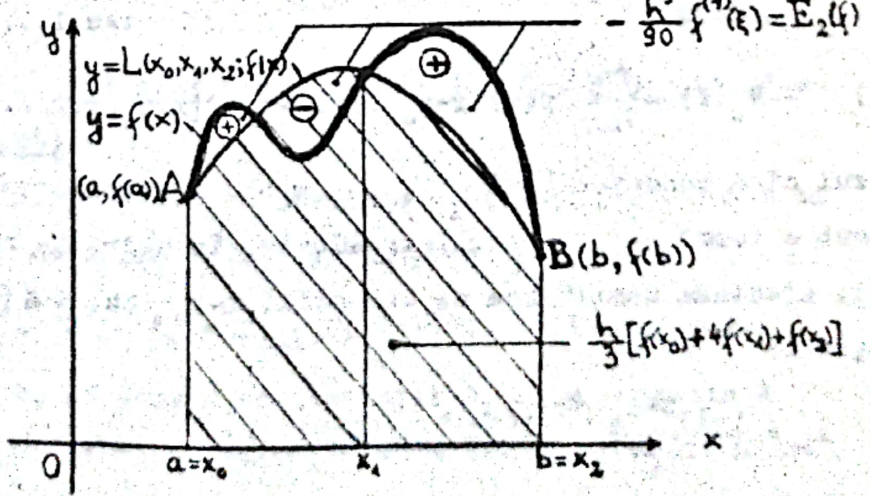


Fig.2.2.

<u>$n = 3$ - formula lui Simpson $\frac{3}{8}$</u>

$$(2.21) \qquad \int_a^b f(x)dx = \frac{3h}{8}\left[f(x_o)+3f(x_1)+3f(x_2)+f(x_3)\right] - \frac{3h^5}{80} f^{(4)}(\xi)$$

unde $x_o = a$, $x_3 = b$, $h = (b-a)/3$ şi $x_o < \xi < x_3$.

Să încheiem aici lista acestor formule.

B. <u>Formule de cuadratură Newton - Côtes deschise</u>

In acest caz relativ la reţeaua de noduri echidistante(2.1) se impun condiţiile $x_o = a+h$ şi $x_n = b-h$; ca urmare intervalul $[a,b]$ este împărţit în $(n+2)$ părţi egale şi $h = (b-a)/(n+2)$. Vom urma întocmai calea parcursă în secţiunea A, modificările uşoare ce vor apărea datorîndu-se poziţiei diferite a nodurilor $x_o, x_1, \ldots, x_n$ în intervalul $[a,b]$. Considerînd formula de interpolare a lui Lagrange pe nodurile $x_o, x_1, \ldots, x_n \in (a,b)$ şi integrînd-o pe intervalul $[a,b]$ se obţine din nou formula de cuadratură (2.7), cu A_i $(i=o,1,\ldots,n)$ şi $E_n(f)$ daţi de (2.8) şi (2.9) (sau (2.10)). Să notăm $x_{-1} = a$ şi $x_{n+1} = b$ (insistăm asupra faptului că punctele x_{-1} şi x_{n+1} nu au rol de noduri !). Atunci formulele (2.8) şi (2.9) pentru coeficienţi şi rest se scriu:

$$(2.22) \qquad A_i = \int_{x_{-1}}^{x_{n+1}} p(x) \frac{(x-x_o)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_o)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}\, dx$$

$$i=o,1,\ldots,n ,$$

şi

$$(2.23) \qquad E_n(f) = \int_{x_{-1}}^{x_{n+1}} p(x)(x-x_o)\ldots(x-x_n)f[x,x_o,x_1,\ldots,x_n]\,dx$$

In cazul cînd ponderea $p(x) = 1$, $x \in [a,b]$, aceste formule se pot pune sub o formă mai simplă. Intr-adevăr, în privinţa relaţiilor (2.22), efectuăm schimbarea de variabilă $x = x_o + th$, $t \in [-1,n+1]$ şi obţinem:

$$A_i = \frac{(-1)^{n-i}h}{i!(n-i)!} \int_{-1}^{n+1} \frac{t(t-1)\ldots(t-n)}{t-i}\, dt , \qquad i = o,1,\ldots,n ,$$

sau

$$(2.24) \qquad A_i = (b-a)\,K_i; \qquad i = 0,1,\dots,n\,,$$

cu

$$(2.25) \qquad K_i = \frac{1}{n+2}\,\frac{(-1)^{n-i}}{i!\,(n-i)!}\int_{-1}^{n+1}\frac{t(t-1)\dots(t-n)}{t-i}\,dt\,, \quad i=0,1,\dots,n$$

In ce privește formula (2.23) pentru rest, calculele pentru aducerea la o formă simplă sînt mult mai anevoioase și pot fi găsite în paragraful următor. Teorema de mai jos strînge rezultatele precedente relativ la coeficienții de cuadratură și indică forma restului potrivit cu paritatea lui n.

Teorema 2.2. Fie $f: [a,b] \to R$, $x_0 = a+h$, $x_n = b-h$, $x_i = x_0 + ih$ $(i = 0,1,\dots,n)$ și $h = \frac{b-a}{n+2}$. Atunci,

(i) dacă n este par și $f \in C^{n+2}[a,b]$, există un punct $\xi \in (a,b)$ astfel încît

$$(2.26) \qquad \int_a^b f(x)dx = (b-a)\sum_{i=0}^{n} K_i f(x_i) + \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!}\int_{-1}^{n+1} t^2(t-1)\dots(t-n)dt\,;$$

(ii) dacă n este impar și $f \in C^{n+1}[a,b]$, există un punct $\xi \in (a,b)$ astfel încît

$$(2.27) \qquad \int_a^b f(x)dx = (b-a)\sum_{i=0}^{n} K_i f(x_i) + \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!}\int_{-1}^{n+1} t(t-1)\dots(t-n)dt\,,$$

unde K_i $(i=0,1,\dots,n)$ sînt coeficienții lui Côtes (de tip deschis) dați de (2.25).

Formulele (2.25) arată că numerele K_i $(i=0,1,\dots,n)$ depind numai de n (și nu de f sau [a,b]) și au proprietățile:
1°. $\sum_{i=0}^{n} K_i = 1$ și 2°. $K_i = K_{n-i}$ $(i=0,1,\dots,n)$ (verificarea acestora se face ca și pentru coeficienții H_i). In consecință se poate alcătui un tabel cu valorile coeficienților K_i funcție de n.

Tabel

Coeficienţii lui Côtes - formulele N-C deschise

n \ K	K_0	K_1	K_2	K_3	K_4	K_5	K_6
o	1						
1	$\frac{1}{2}$	$\frac{1}{2}$					
2	$\frac{2}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$				
3	$\frac{11}{24}$	$\frac{1}{24}$	$\frac{1}{24}$	$\frac{11}{24}$			
4	$\frac{11}{20}$	$-\frac{14}{20}$	$\frac{26}{20}$	$-\frac{14}{20}$	$\frac{11}{20}$		

Observaţii. 1) Formula (2.16) de evaluare a restului (sau a erorii de trunchiere) şi observaţia care succede Teorema 2.1 rămîn valabile şi în cazul formulelor de cuadratură Newton-Côtes deschise.

2) Formulele de cuadratură de tip deschis pot fi utilizate în cazul cînd funcţia ce apare sub integrală nu este definită fie în a, fie în b.

Cele mai simple formule Newton-Côtes de tip deschis, obţinute dînd valori particulare în (2.26) sau (2.27) sînt următoarele:

n = o - formula dreptunghiului:

(2.28) $$\int_a^b f(x)dx = 2h\, f(x_0) + \frac{h^3}{3} f''(\xi)$$

unde $x_0 = \frac{a+b}{2}$, $h = \frac{b-a}{2}$ şi $x_{-1} < \xi < x_1$. Integrala din membrul întîi este aproximată prin aria dreptunghiului mărginit de dreptele $x=a$, $x=b$, $y=0$ şi $y = f(\frac{a+b}{2})$ (fig.2.3).

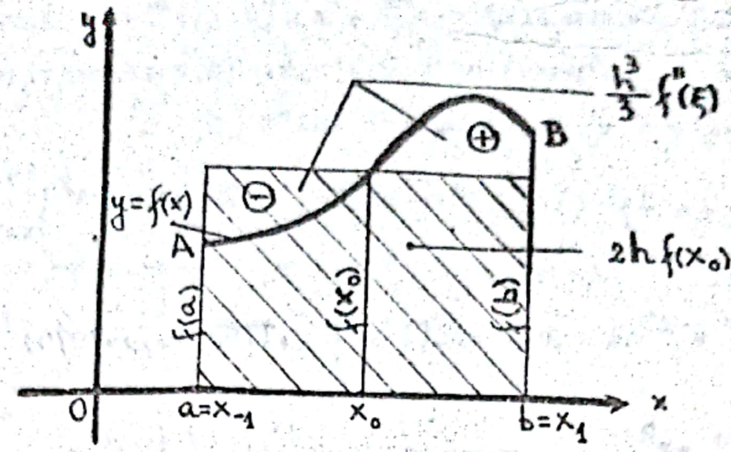


Fig.2.3.

n = 1:

$$(2.29)\qquad \int_a^b f(x)dx = \frac{3h}{2}\left[f(x_0)+f(x_1)\right] + \frac{3h^3}{4}f''(\xi)\ ;$$

unde $x_0 = a+h$, $x_1 = a+2h$, $h = \frac{b-a}{3}$ şi $x_{-1} < \xi < x_2$.

n = 2:

$$(2.30)\qquad \int_a^b f(x)dx = \frac{4h}{3}\left[2f(x_0) - f(x_1) + 2f(x_2)\right] + \frac{14h^5}{45}f^{(4)}(\xi)$$

unde $x_0 = a+h$, $x_2 = b-h$, $h = \frac{b-a}{4}$ şi $x_{-1} < \xi < x_3$.

Observaţie. Alegerea numărului de noduri depinde de acuratețea cu care trebuie calculată integrala funcţiei f cît şi de uşurinţa cu care pot fi calculate valorile funcţiei în noduri. Formulele de cuadratură de tip deschis necesită un număr mai mic de valori ale funcţiei, dar sînt mai puţin precise decît cele de tip închis.

In încheierea paragrafului dăm un exemplu de utilizarea uneia dintre formulele de cuadratură stabilite.

Exemplul 2.1. Să se aproximeze

$$\int_0^1 e^{-x^2}dx$$

folosind formula lui Simpson şi să se evalueze eroarea comisă.

In casul nostru $f(x) = e^{-x^2}$, $x \in [0,1]$, nodurile sînt: $x_0 = 0$, $x_1 = 0{,}5$ şi $x_2 = 1$ şi pasul este $h = 0{,}5$. In conformitate cu formula (2.20), există $\xi \in (0,1)$ astfel încît

$$\int_0^1 e^{-x^2}dx = \frac{0{,}5}{3}(1+4.e^{-0{,}25}+e^{-1}) - \frac{(0{,}5)^5}{90}\left[e^{-x^2}\right]^{(4)}_{x=\xi}$$

Ca urmare

$$\int_0^1 e^{-x^2}dx \approx 0{,}16666(1 + 4.0{,}77880+0{,}367879)$$

adică

$$\int_0^1 e^{-x^2}dx \approx 0{,}747180 \quad .$$

Pe de altă parte, avem succesiv:

$$[e^{-x^2}]' = -2x\, e^{-x^2} \quad ,$$

$$[e^{-x^2}]'' = (4x^2 - 2)e^{-x^2} \quad ,$$

$$[e^{-x^2}]''' = (-8x^3 + 12x)e^{-x^2},$$

$$[e^{-x^2}]^{(4)} = (16x^4 - 48x^2 + 12)e^{-x^2}.$$

şi obţinem, în ce priveşte eroarea,

$$\left| -\frac{(0{,}5)^5}{90}\left[e^{-x^2}\right]^{(4)}_{x=\xi}\right| \leq \frac{1}{32.90}\max_{x\in[0,1]}(|16x^4-48x^2+12|\, e^{-x^2}) \leq$$

$$\leq \frac{4}{32.90}\max_{x\in[0,1]}|4x^4-12x^2+3| = \frac{4.5}{32.90} = 0{,}00694\ \ldots$$

Aşadar, numărul $0{,}747180$ aproximează integrala dată cu o eroare mai mică, în valoare absolută, decît $0{,}007$.

§ 3. Forma restului formulelor Newton-Côtes simple

Să începem cu cîteva chestiuni preliminare. Se numeşte polinom factorial polinomul definit prin

(3.1) $$\pi_n(t) = t(t-1)\ldots(t-n)$$

Se constată cu uşurinţă că π_n are următoarele proprietăţi:

(i) π_n schimbă semnul atunci cînd x trece prin zerourile sale: $0,1,\ldots,n$; într-un interval determinat de două zerouri consecu-

în integrala I_{i+1} ; obținem:

$$I_{i+1} = \int_{i+1}^{i+2} \pi_n(t)dt = \int_i^{i+1} (u+1)u(u-1)\dots(u-n+1)du =$$

$$= \int_i^{i+1} \frac{u+1}{u-n}\, \pi_n(u)du \, .$$

Deoarece $\pi_n(u)$ nu schimbă semnul pe intervalul $[i,i+1]$, putem aplica prima formulă de medie integralei ultime: există un punct $\xi \in (i,i+1)$ astfel încît

$$I_{i+1} = \frac{\xi+1}{\xi-n} I_i \;, \qquad i = 0,1,\dots, \frac{n}{2} -2$$

De aici și din faptul că

$$\left|\frac{\xi+1}{\xi-n}\right| = \frac{\xi+1}{n-\xi} < \frac{(i+1)+1}{n-(i+1)} \leq \frac{\frac{n}{2}}{n-(\frac{n}{2}-1)} = \frac{\frac{n}{2}}{\frac{n}{2}+1} < 1$$

rezultă că

$$|I_{i+1}| < |I_i| \;, \quad i=0,1,\dots,\frac{n}{2} - 2 \;,$$

deci proprietatea 3° este adevărată.

Putem trece la demonstrarea afirmației (iii). Fie, în acest scop, $x \in (0,n)$. Deoarece φ_n satisface (ii), este suficient să ne limităm numai la punctele $x \in (0,\frac{n}{2}]$. Funcția φ_n se poate scrie sub forma:

$$\varphi_n(x) = \sum_{i=0}^{[x]-1} I_i + \int_{[x]}^{x} \pi_n(t)dt$$

In virtutea proprietăților numerelor I_i, avem: $I_0 + I_1 > 0$, $I_2 + I_3 > 0$, etc. Ca urmare, dacă $[x]$ este număr par suma din relația precedentă are un număr par de termeni și deci, este pozitivă. Cum și integrala din membrul drept este pozitivă, rezultă că $\varphi_n(x) > 0$ dacă $[x]$ este pară. Dacă $[x]$ este număr impar, se scrie

$$\varphi_n(x) = \sum_{i=0}^{[x]-2} I_i + (I_{[x]-1} + \int_{[x]}^{x} \pi_n(t)dt)$$

şi se verifică că ambele părţi ce figurează în membrul drept sînt pozitive. Aşadar, $\varphi_n(x) > 0$ oricare ar fi $x \in (0,n)$, q.e.d.

Vom folosi şi următorul rezultat din analiza matematică:

Lema 3.2. Fie $f:[a,b] \to R$ o funcţie continuă arbitrară, punctele $c_1, c_2, \dots, c_k \in [a,b]$ şi scalarii nenegativi $\lambda_1, \lambda_2, \dots, \lambda_k \in R_+$ satisfăcînd la condiţia $\lambda_1 + \lambda_2 + \dots + \lambda_k = 1$. Există un punct $\xi \in [a,b]$ astfel încît

$$f(\xi) = \lambda_1 f(c_1) + \lambda_2 f(c_2) + \dots + \lambda_k f(c_k)$$

Demonstraţie. Avem

$$\min_{[a,b]} f(x) \le \lambda_1 f(c_1) + \lambda_2 f(c_2) + \dots + \lambda_k f(c_k) \le \max_{[a,b]} f(x)$$

Funcţia f fiind continuă, atinge valorile sale extreme şi are proprietatea lui Darboux pe intervalul $[a,b]$. In consecinţă, există $\xi \in [a,b]$ încît are loc egalitatea din enunţ. Q.e.d.

Cu aceste pregătiri puse la punct, să demonstrăm că restul formulei Newton-Côtes simple şi de tip închis este cel indicat de Teorema 2.1 în una sau alta din egalităţile (2.17) sau (2.18), potrivit cu paritatea numărului de noduri.

Introducem funcţia $F:[0,n] \to R$ prin

$$F(t) = f(x_0 + th) \tag{3.3}$$

In virtutea formulei (V,3.14), avem:

$$f[x, x_0, x_1, \dots, x_n] = \frac{1}{h^{n+1}} F[t, 0, 1, \dots, n],$$

unde $x = x_0 + th$.

Să reluăm formula (2.9) pentru rest (cu $p(x) = 1$):

$$E_n(f) = \int_a^b (x-x_0)(x-x_1)\dots(x-x_n) f[x, x_1, \dots, x_n]\, dx$$

şi să efectuăm în integrală schimbarea de variabilă $x = x_0 + th$.

Obţinem:

$$E_n(f) = \int_0^n h^{n+1} t(t-1)\dots(t-n) \frac{1}{h^{n+1}} F[t, 0, 1, \dots, n]\, h\, dt$$

sau

$$E_n(f) = h\int_0^n t(t-1)\dots(t-n)\, F[t,o,1,\dots,n]dt$$

sau

(3.4) $$E_n(f) = h\int_0^n \pi_n(t)\, F[t,o,1,\dots,n]dt$$

Cazul n = par. Prin ipoteză $f\in C^{n+2}[a,b]$ şi deci, graţie formulei (3.3), $F\in C^{n+2}[o,n]$. Integrînd prin părţi, relaţia (3.4) devine:

$$E_n(f)= h[\varphi_n(t)\, F[t,o,1,\dots,n]\Big|_o^n - \int_o^n \varphi_n(t)\, \frac{d}{dt} F[t,o,1,\dots,n]dt]$$

sau, în virtutea Lemei 3.1,

$$E_n(f)= -h\int_o^n \varphi_n(t)\, \frac{d}{dt} F[t,o,1,\dots,n]dt= -h\int_o^n \varphi_n(t) F[t,t,o,1,\dots,n]\, dt .$$

Intrucît φ_n este pozitivă, conform Lemei 3.1, se poate aplica prima teoremă de medie ultimei integrale; se obţine:

$$E_n(f)= -h\, F[\tau,\tau,o,1,\dots,n]\int_o^n \varphi_n(t)dt , \qquad \tau\in(o,n) ,$$

sau, în conformitate cu (V,4.4):

$$E_n(f) = -h\,\frac{F^{(n+2)}(\tau')}{(n+2)!}\int_o^n \varphi_n(t)dt , \qquad \tau'\in(o,n) .$$

Ţinînd seama de (3.3), putem scrie:

$$E_n(f) = -\frac{h^{n+3}}{(n+2)!}\, f^{(n+2)}(\xi)\int_o^n \varphi_n(t)dt, \quad \text{unde}$$

$$\xi = x_o+\tau' h \in(a,b) .$$

Integrînd prin părţi, avem

$$E_n(f)= -\frac{h^{n+3}}{(n+2)!}\, f^{(n+2)}(\xi)\Big[t\,\varphi_n(t)\Big|_o^n - \int_o^n t\,\pi_n(t)dt\Big]$$

adică, utilizînd din nou Lema 3.1 şi avînd în vedere relaţia (3.1),

$$E_n(f)= \frac{h^{n+3}}{(n+2)!}\, f^{(n+2)}(\xi)\int_o^n t^2(t-1)\dots(t-n)dt, \qquad \xi\in(a,b) ,$$

care este tocmai forma (2.17) pentru restul formulei de cuadratură.

Cazul n = impar. In acest caz se presupune că $f \in C^{n+1}[a,b]$ şi ca urmare $F \in C^{n+1}[o,n]$. Expresia (3.4) pentru rest se modifică după cum urmează:

$$E_n(f)= h \int_0^n \pi_{n-1}(t).(t-n)\ F[t,o,1,\dots n]dt =$$

$$= h\Big(\int_0^{n-1} \pi_{n-1}(t)\ (F[t,o,1,\dots,n-1] - F[o,1,\dots,n])\,dt +$$

$$+ \int_{n-1}^{n} \pi_n(t)\ F[t,o,1,\dots,n]\,dt\Big) =$$

$$= E_{n-1}(f) - h\ F[o,1,\dots,n]\int_0^{n-1}\pi_{n-1}(t)dt + h\int_{n-1}^{n}\pi_n(t)F[t,o,1,\dots,n]dt$$

Termenul al doilea este nul, căci

$$(3.5) \qquad \int_0^{n-1}\pi_{n-1}(t)dt = \varphi_{n-1}(n-1) = o\ ,$$

în conformitate cu Lema 3.1, n-1 fiind număr par. In privinţa celui de al treilea termen vom utiliza prima teoremă de medie, $\pi_n(t)$ avînd semn constant pe $[n-1,n]$; vom obţine:

$$E_n(f)= E_{n-1}(f)+ h\ F[\tau,o,1,\dots,n]\int_{n-1}^{n}\pi_n(t)dt\ , \quad \tau \in (n-1,n).$$

Cum pentru $E_{n-1}(f)$ se poate utiliza forma stabilită pentru rest în cazul cînd indicele este par şi cum

$$F[\tau,o,1,\dots,n]= \frac{F^{(n+1)}(\tau')}{(n+1)!} = \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi_2),\ \ \xi_2 = x_o + \tau' h \in (a,b),$$

rezultă că

$$E_n(f)= \frac{h^{n+2}f^{(n+1)}(\xi_1)}{(n+1)!}\int_0^{n-1} t^2(t-1)\dots(t-n+1)dt\ +$$

$$+ \frac{h^{n+2}f^{(n+1)}(\xi_2)}{(n+1)!}\int_{n-1}^{n}\pi_n(t)dt$$

cu $\xi_1, \xi_2 \in (a,b)$. Continuînd a aranja expresia precedentă, avem:

$$E_n(f)=\frac{h^{n+2}}{(n+1)!}\left(f^{(n+1)}(\xi_1)\int_0^{n-1}t(t-1)\dots(t-n+1)(t-n+n)dt+\right.$$

$$\left.+f^{(n+1)}(\xi_2)\int_{n-1}^{n}\pi_n(t)dt\right)=\frac{h^{n+2}}{(n+1)!}\left[f^{(n+1)}(\xi_1)\left(\int_0^{n-1}\pi_n(t)dt+\right.\right.$$

$$\left.\left.+n\int_0^{n-1}\pi_{n-1}(t)dt\right)+f^{(n+1)}(\xi_2)\int_{n-1}^{n}\pi_n(t)dt\right]$$

sau, ţinînd cont de (3.5) ,

$$(3.6)\quad E_n(f)=\frac{h^{n+2}}{(n+1)!}\left(f^{(n+1)}(\xi_1)\int_0^{n-1}\pi_n(t)dt+f^{(n+1)}(\xi_2)\int_{n-1}^{n}\pi_n(t)dt\right),$$

$$\xi_1,\xi_2\in(a,b)\ .$$

Să arătăm acum că integralele din (3.6) au acelaşi semn. Mai întîi

$$\int_{n-1}^{n}\pi_n(t)dt<0\ ,$$

căci $\pi_n(t)$ este negativ pe $[n-1,n]$. Pe de altă parte, în virtutea Lemei 3.1, avem:

$$\int_0^{n-1}\pi_n(t)dt=\varphi_n(n-1)=\varphi_n(1)<0\ .$$

Sîntem în condiţiile care permit aplicarea Lemei 3.2 parantezei din (3.6). Există un punct $\xi\in(a,b)$ astfel încît

$$E_n(f)=\frac{h^{n+2}}{(n+1)!}f^{(n+1)}(\xi)\left(\int_0^{n-1}\pi_n(t)dt+\int_{n-1}^{n}\pi_n(t)dt\right)\ ,$$

deci

$$E_n(f)=\frac{h^{n+2}}{(n+1)!}f^{(n+1)}(\xi)\int_0^{n}\pi_n(t)dt\ ,\qquad \xi\in(a,b)\ ,$$

adică restul sub forma din relaţia (2.18).

In privinţa formelor din Teorema 2.2 pentru resturile formulelor Newton-Côtes simple şi de tip deschis se procedează ca mai sus, considerînd în locul primitivei φ_n a polinomului factorial următoarea:

$$(3.7)\qquad \psi_n(x)=\int_{-1}^{x}\pi_n(t)dt\ ,\qquad x\in\mathbb{R}\ .$$

§ 4. Formule de cuadratură Newton-Côtes iterate

Pentru a aproxima $\int_a^b f(x)dx$ atunci cînd $|b-a|>1$, formulele trapezului, a lui Simpson, a dreptunghiului etc., nu sînt indicate căci restul lor are o valoare absolută relativ mare. Pe de altă parte, dacă s-ar alege un număr de noduri echidistante suficient de mare, atunci polinomul de interpolare pe aceste noduri ar avea gradul destul de mare şi, ca urmare, ar putea să nu fie o aproximaţie bună a funcţiei f. Aceste dificultăţi sînt depăşite prin trecerea de la formulele Newton-Côtes simple la cele iterate, care vor permite obţinerea unei aproximaţii oricît de bună dorim pentru integrala dată. Ideea fundamentală esteurmătoarea: se împarte intervalul de integrare $[a,b]$ într-un număr convenabil de subintervale egale, pe fiecare dintre acestea se aproximează integrala cu ajutorul uneia dintre formulele de cuadratură simple şi se sumează pentru a obţine o aproximaţie a integralei pe întregul interval. Formulele de cuadratură ce se obţin astfel se numesc <u>iterate</u>.

Fie N un număr natural arbitrar. Împărţim intervalul $[a,b]$ în N părţi egale prin punctele

(4.1) $$a_k = a + k\frac{b-a}{N}, \quad k = 0,1,\dots,N .$$

Putem scrie:

(4.2) $$\int_a^b f(x)dx = \sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} f(x)dx .$$

Pe subintervalul $[a_k,a_{k+1}]$, $k = 0,1,\dots,N-1$, considerăm n+1 noduri echidistante date de

(4.3) $$x_{ki} = a_k + i\frac{b-a}{Nn}, \quad i = 0,1,\dots,n,$$

dacă se va utiliza o formulă de tip închis relativ la acest subinterval, sau date de

(4.4) $$x_{ki} = a_k + (i+1)\frac{b-a}{N(n+2)}, \quad i = 0,1,\ldots,n,$$

dacă formula de cuadratură va fi de tip deschis.

Vom prezenta mai întîi formulele Newton-Côtes iterate de tip închis. In acest caz nodurile de cuadratură sînt punctele $x_{ki}=a+(nk+i)\frac{b-a}{nN}$ ($k = 0,1,\ldots,N-1$, $i = 0,1,\ldots,n$) după cum rezultă din (4.1) şi (4.3); printre ele se află şi punctele $a_k (k = 0,1,\ldots,N)$. In egalitatea (4.2) să aplicăm o formulă de tip (2.14) pentru fiecare integrală ce apare în membrul drept; obţinem:

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1}\left\{\frac{b-a}{N}\sum_{i=0}^{n} H_i^n f(x_{ki}) + E_n(f,[a_k,a_{k+1}])\right\}$$

sau

(4.5) $$\int_a^b f(x)dx = \frac{b-a}{N}\sum_{i=0}^{n} H_i^n \sum_{k=0}^{N-1} f(x_{ki}) + \sum_{k=0}^{N-1} E_n(f,[a_k,a_{k+1}]),$$

unde H_i^n $(i = 0,1,\ldots,n)$ sînt coeficienţii lui Côtes depinzînd de n. Pentru a da o formă convenabilă restului din (4.5), vom apela la Teorema 2.1. Distingem două cazuri:

Dacă n = par şi $f \in C^{n+2}[a,b]$, avem:

$$\sum_{k=0}^{N-1} E_n(f,[a_k,a_{k+1}]) = \frac{h^{n+3}}{(n+2)!}\int_0^n t^2(t-1)\ldots(t-n)dt \sum_{k=0}^{N-1} f^{(n+2)}(\xi_k)$$

unde $\xi_k \in (a_k,a_{k+1})$, $k = 0,1,\ldots,N-1$ şi $h = \frac{b-a}{Nn}$. Observăm că, deoarece $f^{(n+2)}$ este continuă, există, în virtutea Lemei 3.2, un punct $\xi \in (a,b)$ astfel încît

$$\frac{1}{N}\sum_{k=0}^{N-1} f^{(n+2)}(\xi_k) = f^{(n+2)}(\xi).$$

Ca urmare, putem scrie:

(4.6) $$\sum_{k=0}^{N-1} E_n(f,[a_k,a_{k+1}]) = \frac{(b-a)h^{n+2}f^{(n+2)}(\xi)}{(n+2)!\,n}\int_0^n t^2(t-1)\ldots(t-n)dt.$$

Dacă n = impar și $f \in C^{n+1}[a,b]$, se utilizează relația(2.18) din Teorema 2.1 și, în mod similar, se ajunge la

$$(4.7)\quad \sum_{k=0}^{N-1} E_n(f,[a_k,a_{k+1}]) = \frac{(b-a)h^{n+1}f^{(n+1)}(\xi)}{(n+1)!\ n}\int_0^h t(t-1)\dots(t-n)dt\ .$$

Grupăm rezultatele date de (4.5), (4.6) și (4.7) într-un singur enunț:

Teorema 4.1. Fie $f:[a,b]\to R$, nodurile $x_{ki} = a+(nk+i)\frac{b-a}{nN}$ $(k=0,1,\dots,N-1,\ i=0,1,\dots,n)$ și $h = \frac{b-a}{Nn}$. Atunci,

(i) dacă n este par și $f \in C^{n+2}[a,b]$, există un punct $\xi \in (a,b)$ astfel încît

$$(4.8)\quad \int_a^b f(x)dx = \frac{b-a}{N}\sum_{i=0}^{n} H_i^n \sum_{k=0}^{N-1} f(x_{ki}) + \frac{(b-a)h^{n+2}f^{(n+2)}(\xi)}{(n+2)!\ n}\int_0^h t^2(t-1)\dots(t-n)dt\ ;$$

(ii) dacă n este impar și $f \in C^{n+1}[a,b]$, există $\xi \in (a,b)$ astfel încît

$$(4.9)\quad \int_a^b f(x)dx = \frac{b-a}{N}\sum_{i=0}^{n} H_i^n \sum_{k=0}^{N-1} f(x_{ki}) + \frac{(b-a)h^{n+1}f^{(n+1)}(\xi)}{(n+1)!\ n}\int_0^h t(t-1)\dots(t-n)dt,$$

H_i^n fiind coeficienții lui Côtes dați de (2.13).

Punem în evidență două cazuri particulare frecvent utilizate.

n = 1 - formula trapezelor relativ la N intervale:

$$(4.10)\quad \int_a^b f(x)dx = \frac{h}{2}\Big[f(a)+2\sum_{k=1}^{N-1} f(x_k)+f(b)\Big] - \frac{(b-a)h^2}{12} f''(\xi),\quad \xi \in (a,b)\ ,$$

unde $h = \frac{b-a}{N}$ și $(x_k)_0^N$ este definit prin $x_k = a+kh$.

Intr-adevăr, să luăm în (4.9) $n = 1$ şi $H_0^1 = H_1^1 = \frac{1}{2}$ (vezi tabelul coeficienţilor Côtes):

$$\int_a^b f(x)dx = \frac{b-a}{2N}\left[\sum_{k=0}^{N-1} f(x_{k0}) + \sum_{k=0}^{N-1} f(x_{k1})\right] - \frac{(b-a)h^2}{12} f''(\xi) =$$

$$= \frac{h}{2}\left[\sum_{k=0}^{N-1} f(a+kh) + \sum_{k=0}^{N-1} f(a+(k+1)h)\right] - \frac{(b-a)h^2}{12} f''(\xi) =$$

$$= \frac{h}{2}\left[f(a) + 2\sum_{k=1}^{N-1} f(a+kh) + f(b)\right] - \frac{(b-a)h^2}{12} f''(\xi) ,$$

adică tocmai formula (4.10).

In mod asemănător se obţine din (4.8) pentru

<u>$n = 2$</u> - <u>formula lui Simpson</u> relativ la N+2 intervale:

$$(4.11)\quad \int_a^b f(x)dx = \frac{h}{3}\left[f(a) + 2\sum_{k=1}^{N-1} f(x_{2k}) + 4\sum_{k=1}^{N} f(x_{2k-1}) + f(b)\right] -$$

$$+ \frac{(b-a)h^4}{180} f^{(4)}(\xi),\quad \xi \in (a,b),$$

unde $h = \frac{b-a}{2N}$ şi $(x_k)_0^{2N}$ este reţeaua nodurilor definită prin $x_k = a+kh$.

In ce priveşte formulele Newton-Côtes iterate, de tip deschis, nodurile sînt punctele $x_{ki} = a + [(n+2)k+(i+1)]\frac{b-a}{(n+2)N}$ $(k=0,1,\dots,N-1;\ i=0,1,\dots,n)$, după cum reiese din (4.1) şi (4.4) Punctele a_k, $k=0,1,\dots,N$, nu se află printre aceste noduri. De această dată, pentru integralele din membrul doi al reţelei (4.2) se utilizează o formulă deschisă de tip (2.27). In urma unor calcule de felul celor de mai sus, sîntem conduşi la

<u>Teorema 4.2</u>. <u>Fie $f:[a,b] \to R$, nodurile $x_{ki} = a + [(n+2)k + (i+1)]\frac{b-a}{(n+2)N}$ $(k=0,1,\dots,N-1;\ i=0,1,\dots,n)$ şi $h = \frac{b-a}{(n+2)N}$. Atunci,</u>

(i) <u>dacă n este par şi $f \in C^{n+2}[a,b]$, există $\xi \in (a,b)$ astfel ca</u>

(4.12) $$\int_a^b f(x)dx = \frac{b-a}{N}\sum_{i=0}^{n} K_i^n \sum_{k=0}^{N-1} f(x_{ki}) +$$

$$+\frac{(b-a)h^{n+2}f^{(n+2)}(\xi)}{(n+2)!\ (n+2)}\int_{-1}^{n+1} t^2(t-1)\dots(t-n)dt\ ;$$

(ii) <u>dacă n este impar și $f\in C^{n+1}[a,b]$, există $\xi\in(a,b)$ astfel ca</u>

(4.13) $$\int_a^b f(x)dx = \frac{b-a}{N}\sum_{i=0}^{n} K_i^n \sum_{k=0}^{N-1} f(x_{ki}) +$$

$$+\frac{(b-a)h^{n+1}f^{(n+1)}(\xi)}{(n+1)!\ (n+2)}\int_{-1}^{n+1} t(t-1)\dots(t-n)dt\ ,$$

K_i^n <u>fiind coeficienții lui Côtes (tip deschis) dați de (2.25)</u>.

Cea mai simplă formulă iterată de tip deschis este următoarea:

<u>n = o</u> - <u>formula dreptunghiurilor</u> relativ la N x 2 intervale :

(4.14) $$\int_a^b f(x)dx = 2h\sum_{k=0}^{N-1} f(x_k) + \frac{(b-a)h^2}{6} f''(\xi),\quad \xi\in(a,b)\ ,$$

unde $h = \frac{b-a}{2N}$ și $(x_k)_o^{N-1}$ este definit prin $x_k = a+(2k+1)h$.

<u>Exemplul 4.1</u>. Să se aproximeze

$$\int_0^1 e^{-x^2}dx$$

cu o eroare mai mică ca 10^{-4}.

Această integrală a fost considerată și în exemplul din § 2. De data aceasta însă vom utiliza formulele de cuadratură iterate, care permit determinarea apriori a numărului N ce asigură că eroarea comisă este $< 10^{-4}$. Valoarea lui N ce trebuiește luată depinde de metodă. Vom avea în vedere metoda dreptunghiurilor, a trapezelor și a lui Simpson iterată (a parabolelor). Resturile acestor metode se citesc din formulele (4.14),(4.10)

şi (4.11), respectiv. Trebuie determinat N astfel încît

$$\left|\frac{(b-a)^3}{6.4\ N^2} f''(\xi)\right| < 10^{-4} \quad \text{(metoda dreptunghiurilor)}$$

sau

$$\left|-\frac{(b-a)^3}{12\ N^2} f''(\xi)\right| < 10^{-4} \quad \text{(metoda trapezelor)}$$

sau

$$\left|-\frac{(b-a)^5}{180.16N^4} f^{(4)}(\xi)\right| < 10^{-4} \quad \text{(metoda lui Simpson)}$$

Este suficient să găsim cel mai mic N natural încît să aibă loc, respectiv,

$$\frac{1}{24\ N^2} \max_{[0,1]} |f''(x)| < 10^{-4}\ , \quad \frac{1}{12\ N^2} \max_{[0,1]} |f''(x)| < 10^{-4} \quad \text{şi}$$

$$\frac{1}{180.16\ N^4} \max_{[0,1]} |f^{(4)}(x)| < 10^{-4}$$

Ţinînd seama că $\max_{[0,1]} |f^{(4)}(x)| \leq 20$ (calculul efectuat în exemplul din § 2) şi că $\max_{[0,1]} |f''(x)| = \max_{[0,1]} |4x^2 - 2|e^{-x^2} \leq 2$, este suficient să găsim un N astfel încît să avem, respectiv,

$$\frac{1}{24\ N^2}\, 2 < 10^{-4}\ , \quad \frac{1}{12\ N^2} .2 < 10^{-4} \quad \text{şi} \quad \frac{1}{180.16\ N^4} .20 < 10^{-4}\ .$$

Aceste inegalităţi sînt echivalente, respectiv, cu

$$12\ N^2 > 10^4\ , \quad 6\ N^2 > 10^4 \quad \text{şi} \quad 9\ N^4 > 625\ ;$$

în consecinţă, integrala dată este exprimată cu o eroare $< 10^{-4}$ dacă luăm N = 29 în cazul metodei dreptunghiurilor, N = 41 în cazul metodei trapezelor şi N = 3 în cazul metodei lui Simpson.

Superioritatea metodei lui Simpson faţă de celelalte două considerate este evidentă şi provine din faptul că restul acestei formule este de ordinul lui h^4. Valoarea aproximativă a integralei, calculată cu metoda lui Simpson pentru N = 3, este

$$\int_0^1 e^{-x^2}dx \approx \frac{1}{18}\left(1+4 \cdot e^{-\frac{1}{36}} + 2\,e^{-\frac{1}{9}} + 4\,e^{-\frac{1}{4}} + 2e^{-\frac{4}{9}} + 4e^{-\frac{25}{36}} + e^{-1}\right)$$

şi deci

$$\int_0^1 e^{-x^2}dx \approx 0{,}746830$$

cu o eroare mai mică ca 10^{-4}.

Vom da ca model algoritmul uneia dintre formulele de cuadratură iterate stabilite mai sus.

<u>Algoritmul formulei lui Simpson.</u> Dată funcţia $f \in C^4[a,b]$, să se aproximeze integrala $I = \int_a^b f(x)dx$ cu o eroare mai mică ca ε.

<u>Pasul 1.</u> Fie N cel mai mic număr natural pentru care

$$\frac{(b-a)^5}{2880} \max_{[a,b]} |f^{(4)}(x)| < \varepsilon .$$

<u>Pasul 2.</u> Fie $h = (b-a)/2N$.

<u>Pasul 3.</u> Fie $x_i = a + ih$, $i=0,1,\ldots,2N$.

<u>Pasul 4.</u> Fie $i = 0$ şi $I = 0$.

<u>Pasul 5.</u> Dacă $i=0$, adună $f(x_i)$ la I şi treci la Pasul 6;
dacă $i=2N$, adună $f(x_i)$ la I şi treci la Pasul 8;
dacă i = par, adună $2f(x_i)$ la I şi treci la Pasul 6;
dacă i = impar, adună $4f(x_i)$ la I şi treci la Pasul 6.

<u>Pasul 6.</u> Adună 1 la i.

<u>Pasul 7.</u> Dacă $i \leq 2N$, treci la Pasul 5.

<u>Pasul 8.</u> Inmulţeşte I cu $\frac{h}{3}$.

<u>Pasul 9.</u> Procedura este completă - I este aproximaţia căutată.

§ 5. <u>Formule de cuadratură de tip Gauss</u>

Ştim că gradul de exactitate m al unei formule de cuadratură interpolative pe n+1 noduri este limitat de $n \leq m \leq 2n+1$ (Teorema 1.1). Am văzut că pentru formulele de cuadratură Newton-Côtes gradul de exactitate atinge limita inferioară n. Acest tip de formule aveau în vedere noduri echidistante. Vom arăta în cele ce urmează că printr-o alegere convenabilă a nodurilor sîntem conduşi la un tip de formule de cuadratură pe n+1 noduri pentru care gradul de exactitate are valoarea maximă 2n+1. Să revedem mai întîi cîteva

cunoştinţe referitoare la polinoamele ortogonale.

Fie (a,b) un interval finit sau nu al dreptei reale şi $L^2[a,b]$ spaţiul Hilbert al funcţiilor $f:(a,b)\longrightarrow R$ de pătrat integrabil (convenim că două funcţii coincid dacă sînt egale aproape peste tot) pe care s-a definit produsul scalar prin

$$<f,g> = \int_a^b p(x)f(x)g(x)dx \quad , \ f,g \in L^2[a,b],$$

unde $p:(a,b)\longrightarrow R$ este ponderea (presupusă pozitivă şi continuă). Să considerăm în acest spaţiu şirul puterilor lui x, adică

(5.1) $$1,\ x,\ x^2,\ldots,x^n,\ldots$$

şi să notăm cu

(5.2) $$X_o(x),\ \ X_1(x),\ldots,X_n(x),\ldots$$

şirul de polinoame obţinute prin ortogonalizare cu ponderea p a şirului (5.1). Aşadar,

(5.3) $$\int_a^b p(x)X_i(x)X_j(x)dx = o \quad , \quad i,j \in N \ , \ i \neq j \ .$$

Este cunoscut faptul că la alegeri ale intervalului (a,b) şi ponderii p corespund, prin ortogonalizarea şirului (5.1), diverse şiruri remarcabile de polinoame ortogonale ce au variate aplicaţii (a se consulta, de exemplu, [19] sau [4]). Cîteva exemple utile: (I) <u>Polinoamele lui Jacobi</u> $P_n^{\alpha,\beta}(x)$, dacă $p(x)=(1-x)^{\alpha}(1+x)^{\beta}$ $\alpha,\beta > -1$, şi se ia intervalul (-1,1); în particular avem:

1°. <u>Polinoamele lui Legendre</u> $P_n(x)$, dacă $\alpha=\beta=o$ (adică $p(x)=1$),

2°. <u>Polinoamele lui Cebîşev de prima specie</u> $T_n(x)$ dacă $\alpha=\beta=-\frac{1}{2}$ (adică $p(x)=\dfrac{1}{1-x^2}$),

3° <u>Polinoamele lui Cebîşev de specia a doua</u> $U_n(x)$ dacă $\alpha=\beta=\frac{1}{2}$ (adică $p(x) = 1-x^2$);

(II) <u>Polinoamele lui Laguerre</u> $L_n(x)$, dacă $p(x)=e^{-x}$ şi se ia intervalul $(o,+\infty)$;

(iii) Polinoamele lui Hermite $H_n(x)$, dacă $p(x)= e^{-x^2}$ şi se ia $(-\infty ,+\infty)$ ca interval.

Revenind la şirul de polinoame ortogonale (5.2), să menţionăm şi faptul că, pentru orice $n \in N$, polinomul $X_n(x)$ are exact n rădăcini distincte situate în intervalul deschis (a,b).

Fie o funcţie $f:(a,b) \longrightarrow R$ şi nodurile $x_0,x_1,\dots,x_n \in (a,b)$.

Teorema 5.1. O condiţie necesară şi suficientă pentru ca formula de cuadratură interpolativă (cu pondere)

$$\int_a^b p(x)f(x)dx = \sum_{i=0}^{n} A_i f(x_i) + E_n(f) \tag{5.4}$$

să aibă gradul de exactitate 2n+1 este ca nodurile $x_0,x_1,\dots,x_n$ să fie rădăcinile polinomului $X_{n+1}(x)$.

Demonstraţie. Necesitatea. Să presupunem că (5.4) este exactă de grad 2n+1. Fie P(x) un polinom oarecare de grad $\leq n$. Atunci $P(x)\,\omega(x)$, unde $\omega(x)=(x-x_0)(x-x_1)\dots(x-x_n)$, este un polinom de grad $\leq 2n+1$ şi, în virtutea ipotezei,avem:

$$\int_a^b p(x)\,P(x)\,\omega(x)dx = \sum_{i=0}^{n} A_i P(x_i)\,\omega(x_i) = 0$$

Ca urmare $\omega(x)$ este ortogonal pe orice polinom de grad cel mult n şi, deci, coincide pînă la un factor constant cu $X_{n+1}(x)$: $\omega(x) = c\,X_{n+1}(x)$. De aici rezultă că nodurile $x_0,x_1,\dots,x_n$ sînt rădăcinile lui $X_{n+1}(x)$.

Suficienţa. Presupunem că $X_{n+1}(x)$ are ca rădăcini nodurile $x_0,x_1,\dots,x_n$. Să verificăm că formula (5.4) este exactă pentru orice polinom Q(x) de grad $\leq 2n+1$. Impărţind Q(x) la $X_{n+1}(x)$, avem:

$$Q(x) = X_{n+1}(x)S(x)+ R(x) \quad ,$$

unde S(x) şi R(x) sînt polinoame ce au gradele cel mult n. Prin integrare obţinem:

$$\int_a^b p(x)Q(x)dx = \int_a^b p(x)X_{n+1}(x)S(x)dx + \int_a^b p(x)R(x)dx \quad .$$

Prima integrală din membrul drept este nulă, deoarece $X_{n+1}(x)$ este ortogonal cu orice polinom de grad $\leq n$. Relativ la a doua integrală utilizăm faptul că orice formulă de cuadratură interpolativă are gradul de exactitate $\geqslant n$. Putem scrie:

$$\int_a^b p(x)\ Q(x)dx = \sum_{i=0}^{n} A_i\ R(x_i) \quad .$$

Cum însă $Q(x_i) = X_{n+1}(x_i)S(x_i) + R(x_i) = R(x_i)$, rezultă că

$$\int_a^b p(x)\ Q(x)dx = \sum_{i=1}^{n} A_i\ Q(x_i) \ ,$$

adică (5.4) este exactă pentru orice polinom de grad $\leqq 2n+1$. Gradul de exactitate al formulei (5.4) este, deci, $\geqslant 2n+1$. Cum acest număr nu poate fi depășit rezultă că (5.4) are gradul de exactitate egal cu $2n+1$. Q.e.d.

Formulele de cuadratură interpolative de forma (5.4), cu nodurile $x_0, x_1, \dots, x_n$ tocmai rădăcinile polinomului ortogonal $X_{n+1}(x)$ se numesc <u>formule de cuadratură ale lui Gauss.</u> Pentru a obține expresiile coeficienților și restului acestor formule se poate proceda pe o cale similară cu cea urmată în cazul formulelor Newton-Côtes: scriem formula de interpolare a lui Hermite pe rădăcinile polinomului $X_{n+1}(x)$ ca noduri și o integrăm pe intervalul (a,b) (după ce am înmulțind-o mai întîi cu ponderea $p(x)$).

<u>Teorema 5.2.</u> <u>Fie $f \in C^{2n+2}(a,b)$ și $x_0, x_1, \dots, x_n$ rădăcinile polinomului $X_{n+1}(x)$ (despre care presupunem că are coeficientul director 1). Atunci există un punct $\xi \in (a,b)$ astfel încît</u>

$$(5.5) \qquad \int_a^b p(x)f(x)dx = \sum_{i=0}^{n} A_i f(x_i) + \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b p(x)\left[X_{n+1}(x)\right]^2 dx,$$

unde

$$A_i = \frac{1}{\left[X'_{n+1}(x_i)\right]^2}\int_a^b p(x)\left[\frac{X_{n+1}(x)}{x-x_i}\right]^2 dx\ , \qquad i=o,1,\dots,n \tag{5.6}$$

Demonstrație. Formula de interpolare a lui Hermite pe rădăcinile $x_o,x_1,\dots,x_n$ ale polinomului $X_{n+1}(x)$ este:

$$f(x) = H(x_o,2;\ x_1,2;\dots;x_n,2;\ f|x) + R_{2n+1}(x) \tag{5.7}$$

unde

$$H(x_o,2;\dots;x_n,2;\ f|x) = \sum_{i=o}^{n}\left[1-2L'_i(x_i)(x-x_i)\right]\left[L_i(x)\right]^2 f(x_i) + \sum_{i=o}^{n}(x-x_i)\left[L_i(x)\right]^2 f'(x_i) \tag{5.8}$$

cu

$$L_i(x) = \frac{X_{n+1}(x)}{(x-x_i)X'_{n+1}(x_i)}\ , \qquad i=o,1,\dots,n\ , \tag{5.9}$$

și restul dat de

$$R_{2n+1}(x) = \frac{f^{(2n+2)}(\xi')}{(2n+2)!}\left[X_{n+1}(x)\right]^2\ , \qquad \xi'\in(a,b)\ . \tag{5.10}$$

Inmulțind formula de interpolare (5.7) cu $p(x)$ și integrînd pe (a,b), obținem:

$$\int_a^b p(x)f(x)dx = \sum_{i=o}^{n} A_i f(x_i) + \sum_{i=o}^{n} B_i f'(x_i) + \int_a^b p(x)R_{2n+1}(x)dx, \tag{5.11}$$

coeficienții A_i și B_i, $i=o,1,\dots,n$, fiind dați de

$$A_i = \int_a^b p(x)\left[1-2\,L'_i(x_i)(x-x_i)\right]\left[L_i(x)\right]^2 dx\ , \tag{5.12}$$

și

$$B_i = \int_a^b p(x)(x-x_i)\left[L_i(x)\right]^2 dx\ . \tag{5.13}$$

Calculăm mai întîi coeficienții A_i. Ținînd seama de (5.9), vom avea:

$$A_i = \int_a^b p(x)\left[L_i(x)\right]^2 dx - 2L'_i(x_i)\int_a^b p(x)(x-x_i)\left[L_i(x)\right]^2 dx =$$

$$= \frac{1}{\left[X'_{n+1}(x_i)\right]^2}\left\{\int_a^b p(x)\left[\frac{X_{n+1}(x)}{x-x_i}\right]^2 dx - 2L'_i(x_i)\int_a^b p(x)\,\frac{X_{n+1}(x)}{x-x_i}X_{n+1}(x)dx\right\}.$$

Ultima integrală este nulă, deoarece $X_{n+1}(x)$ este ortogonal pe polinomul $\frac{X_{n+1}(x)}{x-x_i}$ de grad n. Ca urmare

$$A_i = \frac{1}{\left[X'_{n+1}(x_i)\right]^2}\int_a^b p(x)\left[\frac{X_{n+1}(x)}{x-x_i}\right]^2 dx \quad , i = 0,1,\dots,n ,$$

adică tocmai formulele (5.6). In ce priveşte coeficienţii B_i, rezultă uşor din (5.13) şi (5.9) că ei sînt nuli:

$$B_i = \frac{1}{\left[X'_{n+1}(x_i)\right]^2}\int_a^b p(x)\,\frac{X_{n+1}(x)}{x-x_i}\,X_n(x)dx = 0 \ , \quad i=0,1,\dots,n .$$

Relativ la restul formulei de cuadratură avem:

$$E_n(f)= \int_a^b p(x)R_{2n+1}(x)dx = \frac{1}{(2n+2)!}\int_a^b p(x)\left[X_{n+1}(x)\right]^2 f^{(2n+2)}(\xi')dx ,$$

unde $\xi'=\xi'(x) \in (a,b)$. Deoarece funcţia $p(x)\left[X_{n+1}(x)\right]^2$ este pozitivă, putem aplica prima teoremă de medie şi scrie:

$$(5.14) \qquad E_n(f)= \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b p(x)\left[X_{n+1}(x)\right]^2 dx \ ,$$

unde $\xi \in (a,b)$. Introducînd în (5.11) valorile găsite pentru coeficienţi şi expresia pentru rest se obţine formula de cuadratură (5.5). Q.e.d.

Observaţii. 1) Formulele (5.6) arată că valorile coeficienţilor formulei de cuadratură a lui Gauss sînt pozitive şi că ele rămîn valabile şi în cazul cînd $X_{n+1}(x)$ nu are coeficientul director egal cu 1.

2) Formulele de cuadratură ale lui Gauss sînt de tip deschis.

§ 6. Formule de cuadratură Gauss-Legendre, Gauss-Laguerre și Gauss-Hermite

Să considerăm $p(x) = 1$, $x \in (a,b)$, $a = -1$ și $b = +1$. Ca urmare, șirul de polinoame ortogonale (5.2) va fi șirul polinoamelor lui Legendre $P_n(x)$, $n \in N$. Formula lui Rodrigues pentru aceste polinoame este

(6.1) $$P_n(x) = \frac{1}{2^n \cdot n!} \frac{d^n}{dx^n} [(x^2-1)^n] \quad , \qquad n \in N \quad .$$

Au loc relațiile: $P_n(-x) = (-1)^n P_n(x)$ și $P_n(1) = 1$, $n \in N$.

Primele polinoame Legendre sînt: $P_0(x)=1$, $P_1(x)=x$, $P_2(x)= = \frac{1}{2}(3x^2-1)$, $P_3(x)= \frac{1}{2}(5x^3-3x)$, etc. Rădăcinile polinomului $P_{n+1}(x)$ sînt simple, aparțin intervalului $(-1,1)$ și, ținînd seama de proprietatea de paritate de mai sus, sînt simetrice față de origine, adică avem $x_k = -x_{n-k}$, $k=0,1,\dots,n$ (am presupus că aceste rădăcini sînt numerotate în ordine crescătoare: $x_0 < x_1 < \dots < x_n$). Alte proprietăți remarcabile ale acestor polinoame pot fi găsite în [4].

In condițiile particulare în care sîntem plasați, coeficienții și restul formulei de cuadratură (5.5) pot căpăta expresii mai simple. Pornind de la relațiile (5.6) și efectuînd o integrare prin părți, obținem:

$$[P'_{n+1}(x_i)]^2 A_i = \int_{-1}^{+1} \left[\frac{P_{n+1}(x)}{x-x_i}\right]^2 dx = -\frac{[P_{n+1}(x)]^2}{x-x_i}\Bigg|_{-1}^{+1} + 2\int_{-1}^{+1} \frac{P_{n+1}(x)P'_{n+1}(x)}{x-x_i}dx$$

$$= -\left(\frac{1}{1-x_i} + \frac{1}{1+x_i}\right) + 2\int_{-1}^{+1} \frac{P_{n+1}(x)P'_{n+1}(x)}{x-x_i}dx \quad .$$

Atunci, ținînd seama că formula de cuadratură este exactă pentru polinoamele de grad $\leq 2n+1$, avem :

$$\left[P'_{n+1}(x_i)\right]^2 A_i = \frac{2}{x_i^2-1} + 2\sum_{k=0}^{n} A_k\left[\frac{P_{n+1}(x)P'_{n+1}(x)}{x-x_i}\right]_{x=x_k} =$$

$$= \frac{2}{x_i^2-1} + 2\sum_{k=0}^{n} A_k\left[\frac{P_{n+1}(x)}{x-x_i}\right]_{x=x_k} \cdot P'_{n+1}(x_k) =$$

$$= \frac{2}{x_i^2-1} + 2\, A_i\left[\frac{P_{n+1}(x)}{x-x_i}\right]_{x=x_i} \cdot P'_{n+1}(x_i) .$$

Dar $P'_{n+1}(x) = \sum\limits_{j=0}^{n} \frac{P_{n+1}(x)}{x-x_j}$ şi deci $P'_{n+1}(x_i) = \left[\frac{P_{n+1}(x)}{x-x_i}\right]_{x=x_i}$;

ca urmare

$$\left[P'_{n+1}(x_i)\right]^2 A_i = \frac{2}{x_i^2-1} + 2\, A_i\left[P'_{n+1}(x_i)\right]^2$$

sau

$$A_i = \frac{2}{(1-x_i^2)\left[P'_{n+1}(x_i)\right]^2} \quad , \quad i = 0,1,\dots,n . \tag{6.2}$$

<u>Observaţie</u>. Coeficienţii formulei de cuadratură Gauss-Legendre sînt pozitivi şi au proprietatea de simetrie

$$A_i = A_{n-i} \qquad , \; i = 0,1,\dots,n . \tag{6.3}$$

Intr-adevăr, afirmaţia rezultă din (6.2) şi faptul că $x_i = -x_{n-i}$, $i = 0,1,\dots,n$, şi $P'_{n+1}(-x) = (-1)^n P'_{n+1}(x)$ (ultima egalitate obţinîndu-se prin derivarea relaţiei $P_{n+1}(-x) = (-1)^{n+1}P_{n+1}(x)$).

In ce priveşte restul, să ţinem seama de (6.1) şi să înlocuim în (5.14) $X_{n+1}(x)$ cu expresia obţinută graţie formulei

$$X_n(x) = \frac{2^n(n!)^2}{(2n)!} P_n(x) = \frac{n!}{(2n)!} \frac{d^n}{dx^n}\left[(x^2-1)^n\right]$$

(prezenţa factorului numeric din membrul doi face ca $X_n(x)$ să aibă coeficientul director 1); vom avea

$$E_n(f) = \left[\frac{2^{n+1}[(n+1)!]^2}{(2n+2)!}\right]^2 \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^{+1}\left[P_{n+1}(x)\right]^2 dx .$$

Se ştie însă că

(6.4) $$\|P_n\|^2 = \int_{-1}^{+1} [P_n(x)]^2 dx = \frac{2}{2n+1}$$ ([4], p.259) .

Ca urmare

(6.5) $$E_n(f) = \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi), \quad \xi \in (-1,1).$$

Aşadar, am demonstrat următorul rezultat

Teorema 6.1. Dacă $f \in C^{2n+2}[-1,1]$ şi $x_o < x_1 < \ldots < x_n$ sînt rădăcinile polinomului $P_{n+1}(x)$, atunci există un punct $\xi \in (-1,1)$ astfel încît are loc următoarea formulă de cuadratură Gauss-Legendre:

(6.6) $$\int_{-1}^{+1} f(x)dx = \sum_{i=o}^{n} A_i f(x_i) + \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi),$$

valorile coeficienţilor A_i fiind date de formulele (6.2).

Un inconvenient al formulei de cuadratură Gauss-Legendre este faptul că rădăcinile $x_o, x_1, \ldots, x_n$ ale polinomului $P_{n+1}(x)$, precum şi coeficienţii $A_o, A_1, \ldots, A_n$ sînt în general numere iraţionale, ceea ce conduce la calcule complicate. Acest neajuns este compensat în parte de o precizie mare a aproximării pentru un număr relativ mic de valori ale funcţiei.

Tabelul nodurilor şi coeficienţilor f. de c. Gauss-Legendre

n	i	x_i	A_i
o	o	o	2
1	o;1	∓ o,57735o27	1
2	o;2	∓ o,77459667	$\frac{5}{9}$ = o,55555556
	1	o	$\frac{8}{9}$ = o,88888889
3	o;3	∓ o,86113631	o,34785484
	1;2	∓ o,33998104	o,65214516
4	o;4	∓ o,9o617985	o,23692688
	1;3	∓ o,53846931	o,47862868
	2	o	o,56888889

Pentru a calcula

$$\int_a^b f(x)dx$$

utilizînd formulele de cuadratură Gauss-Legendre efectuăm schimbarea de variabilă

(6.7) $$x = \frac{a+b}{2} + \frac{b-a}{2}t \quad , \; t\in[-1,1] \; ;$$

obţinem

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right)dt \quad .$$

Conform formulei de cuadratură amintite, avem:

(6.8) $$\int_a^b f(x)dx = \frac{b-a}{2}\sum_{i=0}^{n} A_i f\left(\frac{a+b}{2} + \frac{b-a}{2}t_i\right) + E_n(f,[a,b]) \; ,$$

unde $t_i\in[-1,1]$, $i=0,1,\ldots,n$, sînt rădăcinile polinomului Legendre $P_{n+1}(t)$. Coeficienţii A_i sînt daţi tot de formulele (6.2), iar pentru rest avem, ţinînd seama de (6.5),

$$E_n(f,[a,b]) = \frac{b-a}{2}E\left(f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right), [-1,1]\right) =$$

$$= \frac{b-a}{2}\,\frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}\left[f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right)\right]^{(2n+2)}_{t=\tau} ,$$

$$\tau\in(-1,1),$$

adică

(6.9) $$E_n(f,[a,b]) = \frac{(b-a)^{2n+3}[(n+1)!]^4}{(2n+3)\,[(2n+2)!]^3} f^{(2n+3)}(\xi) \; ,$$

unde $\xi = \frac{a+b}{2} + \frac{b-a}{2}\tau\in(a,b)$. Sîntem conduşi astfel la

<u>Teorema 6.2. Dacă $f\in C^{2n+2}[a,b]$ şi $x_i = \frac{a+b}{2} + \frac{b-a}{2}t_i$, $i=0,1,\ldots,n$, unde t_i, $i=0,1,\ldots,n$, sînt zerourile polinomului Legendre $P_{n+1}(t)$, atunci există un punct $\xi\in(a,b)$ astfel încît are loc</u>

$$(6.10)\qquad \int_a^b f(x)dx = \frac{b-a}{2}\sum_{i=0}^{n} A_i f(x_i) + \frac{(b-a)^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi)\ ,$$

coeficienţii de cuadratură A_i avînd expresiile (6.2).

Nodurile $x_i \in (a,b)$, $i=o,1,\dots,n$, sînt simetrice faţă de mijlocul $\frac{a+b}{2}$ al acestui interval întrucît sînt imaginile prin (6.7) ale rădăcinilor polinomului lui Legendre $P_{n+1}(t)$ şi $\frac{a+b}{2}$ corespunde la $t=o$. Altfel spus, punctele x_i sînt dispuse în (a,b) „asemănător" cu punctele t_i în intervalul $(-1,1)$.

Să trecem la stabilirea formulei Gauss-Legendre iterate, care este recomandabilă în cazul cînd $|b-a|>1$ (în acest caz expresia restului formulei simple arată că valoarea absolută a acestuia poate fi relativ mare). Impărţim intervalul $[a,b]$ în N părţi egale prin punctele

$$a_k = a + k\,\frac{b-a}{N}\ , \qquad k = o,1,\dots,N\ .$$

Pe fiecare subinterval $[a_k,a_{k+1}]$, $k=o,1,\dots,N-1$, considerăm următoarele $n+1$ noduri:

$$(6.11)\qquad x_{ki} = \frac{a_k+a_{k+1}}{2} + \frac{b-a}{2N}\,t_i\ , \qquad i = o,1,\dots,n\ ,$$

unde t_i sînt rădăcinile polinomului lui Legendre $P_{n+1}(t)$. In virtutea formulei de cuadratură Gauss-Legendre simple, aplicată fiecărui interval $[a_k,a_{k+1}]$, avem:

$$\int_a^b f(x)dx = \sum_{k=o}^{N-1}\int_{a_k}^{a_{k+1}} f(x)dx = \frac{b-a}{2N}\sum_{k=o}^{N-1}\sum_{i=1}^{n} A_i f(x_{ki}) + \sum_{k=o}^{N-1} E_n(f,[a_k,a_{k+1}])$$

sau, avînd în vedere (6.9),

$$\int_a^b f(x)dx = \frac{b-a}{2N}\sum_{i=0}^{n} A_i \sum_{k=0}^{N-1} f(x_{ki}) +$$

$$+ \left(\frac{b-a}{N}\right)^{2n+3} \frac{[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} \sum_{k=0}^{N-1} f^{(2n+2)}(\xi_k) ,$$

unde $\xi_k \in (a_k, a_{k+1})$, $k=0,1,\dots,N-1$. In virtutea Lemei 3.2, există $\xi \in (a,b)$ aşa că

$$f^{(2n+2)}(\xi) = \frac{1}{N}\sum_{k=0}^{N-1} f^{(2n+2)}(\xi_k)$$

şi ca urmare obţinem

Teorema 6.3. Dacă $f \in C^{2n+2}[a,b]$ şi t_i, $i=0,1,\dots,n$, sînt rădăcinile polinomului Legendre $P_{n+1}(t)$, atunci există $\xi \in (a,b)$ aşa încît are loc formula de cuadratură Gauss-Legendre iterată relativ la $N \times (n+1)$ intervale:

$$(6.12) \qquad \int_a^b f(x)dx = \frac{b-a}{2N}\sum_{i=0}^{n} A_i \sum_{k=0}^{N-1} f(x_{ki}) +$$

$$+ \frac{(b-a)^{2n+3}[(n+1)!]^4}{N^{2n+2}(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi) ,$$

unde x_{ki} ($k=0,1,\dots,N-1$, $i=0,1,\dots,n$) sînt daţi de (6.11) iar coeficienţii A_i ($i=0,1,\dots,n$) de relaţiile (6.2).

Observaţie. Atît formulele de cuadratură Gauss-Legendre simple, cît şi cele iterate au gradul de exactitate $2n+1$ (ceea ce se poate vedea direct din (6.10) şi (6.12)).

In cazul particular $n=0$, formula (6.12) devine:

$$(6.13) \qquad \int_a^b f(x)dx = \frac{b-a}{N}\sum_{k=0}^{N-1} f\left(\frac{a_k+a_{k+1}}{2}\right) + \frac{(b-a)^3}{24 N^2} f''(\xi), \quad \xi \in (a,b)$$

(am avut în vedere că $t_0=0$ şi $A_0=2$, ceea ce rezultă direct sau din tabelul dat anterior). Aceasta este tocmai formula dreptunghiurilor.

Expresiile resturilor formulei (6.12) pentru primele trei valori ale lui n sînt următoarele:

(6.14) $$\frac{(b-a)^3}{24N^2} f''(\xi)\ , \quad \text{dacă } n = 0 \quad ;$$

(6.15) $$\frac{(b-a)^5}{4320N^4} f^{(4)}(\xi)\ , \quad \text{dacă } n = 1;$$

(6.16) $$\frac{(b-a)^7}{2016000\ N^6} f^{(6)}(\xi), \quad \text{dacă } n = 2.$$

Să comparăm formula Gauss-Legendre iterată pentru n=1 cu formula Simpson iterată relativ la N x 2 intervale. Ambele au gradul de exactitate egal cu trei. Restul primei formule este dat de (6.15), iar restul celei din urmă de expresia (a se vedea (4.11)!):

(6.17) $$-\frac{(b-a)^5}{2880\ N^4} f^{(4)}(\xi)\ .$$

Observăm că restul formulei Gauss-Legendre are coeficientul numeric de 1,5 ori mai mic în valoare absolută decît cel al restului formulei Simpson. Pe de altă parte, numărul de valori ale funcţiei f ce trebuiesc calculate este 2N în cazul primei formule şi 2N+1 în cazul celei de a doua. Cum nodurile formulei Simpson iterate sînt echidistante, calculul celor 2N+1 valori cerute de formulă va fi mai comod. Tragem concluzia că va trebui să preferăm formula de cuadratură a lui Simpson iterată şi că formula Gauss-Legendre iterată se va utiliza numai pentru n ≥ 2.

Exemplul 6.1. Utilizînd formula de cuadratură Gauss-Legendre iterată cu n = 1, să se aproximeze

$$\int_0^1 e^{-x^2}\,dx$$

cu o eroare mai mică ca 10^{-4}.

In Exemplul 4.1 acelaşi exerciţiu a fost abordat cu metoda Simpson iterată. Vom utiliza unele calcule efectuate cu acest

prilej. In cazul nostru, conform formulei (6.15), determinăm N din condiţia

$$\frac{1}{4320\, N^4} \max_{[0,1]} |f^{(4)}(x)| < 10^{-4}$$

sau, deoarece $\max_{[0,1]} |f^{(4)}(x)| \leq 20$, impunînd condiţia mai tare

$$\frac{20}{4320\, N^4} < 10^{-4},$$

care ne conduce la $N = 3$. Aceeaşi valoare pentru N a fost găsită şi în cadrul metodei Simpson; vom vedea în continuare că lucrul cu metoda propusă este mai anevoios. Luînd $n=1$ şi $N=3$ în (6.12), avem

$$\int_0^1 e^{-x^2} dx \approx \frac{1}{6} \sum_{i=0}^{n} A_i \sum_{k=0}^{2} f(x_{ki}) \quad .$$

Dar la $n=1$ tabelul relativ la formula Gauss-Legendre indică $A_0 = A_1 = 1$ şi $t_0 = -t_1 = -0,57735$. Cum (6.11) conduce la $x_{ki} = \frac{1}{3}(k + \frac{1}{2}) + \frac{1}{6} t_i$ ($k = 0,1,2$ şi $i = 0,1$), obţinem: $x_{00} = \frac{1}{6} + \frac{1}{6} t_0$, $x_{01} = \frac{1}{6} + \frac{1}{6} t_1$, $x_{10} = \frac{1}{2} + \frac{1}{6} t_0$, $x_{11} = \frac{1}{2} + \frac{1}{6} t_1$, $x_{20} = \frac{5}{6} + \frac{1}{6} t_0$, $x_{21} = \frac{5}{6} + \frac{1}{6} t_1$, adică $x_{00} = 0,070441$, $x_{01} = 0,262891$, $x_{10} = 0,403775$, $x_{11} = 0,596224$, $x_{20} = 0,737108$ şi $x_{21} = 0,929558$. Atunci

$$\int_0^1 e^{-x^2} dx \approx 0,166667 \left[\exp(-0,070441^2) + \right.$$
$$+ \exp(-0,262891^2) + \exp(-0,403775^2) + \exp(-0,596224^2) +$$
$$\left. + \exp(-0,737108^2) + \exp(-0,929558^2) \right],$$

de unde rezultatul final

$$\int_0^1 e^{-x^2} dx \approx 0,746821 \quad .$$

x

x x

In continuare vom prezenta sumar formulele de cuadratură Gauss-Laguerre şi Gauss-Hermite. Calea de obţinere a acestora reproduce pe cea urmată în cazul formulelor Gauss-Legendre.

Fie a = o, b = +∞ şi $p(x) = e^{-x}$ dacă $x \in (o, \infty)$. Prin ortogonalizarea şirului (5.1) se obţine şirul polinoamelor lui Laguerre $L_n(x)$. Amintim formula lui Rodrigues pentru aceste polinoame:

$$L_n(x) = \frac{1}{n!} e^x \frac{d^n}{dx^n} (e^{-x} x^n)$$

şi faptul că $\|L_n\| = 1$, $n \in N$. Cu aceste elemente se poate obţine ca mai sus formula de cuadratură Gauss-Laguerre

(6-18) $$\int_o^\infty e^{-x} f(x)dx = \sum_{i=o}^{n} \frac{f(x_i)}{x_i \left[L'_{n+1}(x_i)\right]^2} + \frac{1}{(2n+2)!} f^{(2n+2)}(\xi),$$

$\xi > o$,

unde $x_o, x_1, \dots, x_n$ sînt rădăcinile polinomului $L_{n+1}(x)$.

Tabelul nodurilor şi coeficienţilor f. de c. Gauss-Laguerre

n	i	x_i	A_i
o	o	1	1
1	o	o,5857864	o,8535533
	1	3,4142135	o,1464466
2	o	o,4157745	o,711o93o
	1	2,29428o3	o,2785177
	2	6,289945o	o,o1o3892
3	o	o,3225476	o,6o31541
	1	1,7457611	o,3574186
	2	4,53662o2	o,o388879
	3	9,395o7o9	o,ooo5392

Considerăm acum a = −∞, b = +∞ şi ponderea $p(x)=e^{-x^2}$, x (−∞,+∞). Polinoamele ortogonale asociate acestora sînt polinoamele lui Hermite $H_n(x)$. Formula lui Rodrigues este

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

şi $\|H_n\|^2 = 2^n n! \sqrt{\pi}$. Procedind ca mai sus, se obţine <u>formula de cuadratură Gauss-Hermite</u>:

$$\text{(6.19)} \quad \int_{-\infty}^{+\infty} e^{-x^2} f(x)dx = 2^{n+2}(n+1)!\sqrt{\pi} \sum_{i=0}^{n} \frac{f(x_i)}{\left[H'_{n+1}(x_i)\right]^2} +$$

$$+ \frac{(n+1)!\sqrt{\pi}}{2^{n+1}(2n+2)!} f^{(2n+2)}(\xi) ,$$

unde $x_0 < x_1 < \dots < x_n$ sînt rădăcinile polinomului $H_{n+1}(x)$.

Tabelul nodurilor şi coeficienţilor f. de c. Gauss-Hermite

n	i	x_i	A_i
o	o	o	1,7724538
1	o;1	∓ o,7o71o67	o,8862269
2	o;2	∓ 1,2247448	o,2954o89
	1	o	1,1816359
3	o;3	∓ 1,65o68o1	o,o813128
	1;2	∓ o,5246476	o,8o4914o

<u>Observaţie.</u> Am văzut că formulele de cuadratură ale lui Gauss sînt de tip deschis. Ca urmare, în obţinerea formulelor iterate nici una dintre valorile funcţiei calculate pentru un subinterval nu este folosită la subintervalul următor. Aşa numitele <u>formule de cuadratură ale lui Lobatto</u> sînt formule de grad maxim de exactitate şi au extremităţile intervalului (sau intervalelor) printre noduri).

§ 7. <u>Observaţii asupra erorilor şi convergenţei formulelor de cuadratură</u>

Se pune problema de a evalua eroarea totală a unei formule de cuadratură. Reluăm o astfel de formulă sub formă generală:

$$\text{(7.1)} \quad \int_a^b p(x)f(x)dx = \sum_{i=0}^{n} A_i f(x_i) + R_n(f) \quad -$$

Eroarea totală a formulei este suma dintre eroarea de metodă (eroarea de trunchiere) dată de restul formulei şi cea provenită prin acumularea erorilor de rotunjire.

Să ne limităm la formulele de cuadratură pe noduri echidistante (Newton-Côtes). Fie h pasul reţelei de noduri (egale cu $\frac{b-a}{n}$ sau $\frac{b-a}{n+2}$ după cum formula este de tip închis sau deschis). Precizia formulei de cuadratură este determinată mai ales de ordinul restului $E_n(f)$ în raport cu pasul h. Astfel, în conformitate cu (4.10) şi (4.14), restul formulelor de cuadratură ale trapezelor şi dreptunghiurilor sînt de ordinul doi în raport cu h (altfel spus, sînt de ordinul lui h^2). In ce priveşte formula de cuadratură a lui Simpson (4.11), arată că restul ei este de ordinul lui h^4. Aceasta explică faptul că pentru funcţii de clasă $C^4[a,b]$ formula lui Simpson dă, pentru valori suficient de mici ale lui h, o precizie a aproximării mult mai bună decît celelalte două. Deseori evaluarea erorii de trunchiere a unei formule de cuadratură este anevoioasă pe baza expresiei pentru rest. Practic se poate proceda astfel: se utilizează calculul dublu pentru paşii h şi $2h$ şi se consideră că zecimalele ce coincid sînt corecte.

Să trecem acum la cîteva observaţii relativ la erorile de rotunjire inerente utilizării unei formule de cuadratură (presupunem în continuare nodurile echidistante). Să considerăm că valorile $f(x_i)$ ale funcţiei în noduri au fost calculate aproximativ: $f(x_i)= \bar{f}(x_i) + e_i$, $i=0,1,\dots,n$, unde e_i notează eroarea de rotunjire asociată cu valoarea calculată $\bar{f}(x_i)$. In cazul nostru $A_i= (b-a) H_i$ sau $A_i= (b-a) K_i$, $i=0,1,\dots,n$ şi ca urmare aceşti coeficienţi sînt constante exacte. Atunci

$$(7.2) \qquad \int_a^b p(x)f(x)dx \approx (b-a) \sum_{i=0}^{n} H_i\bar{f}(x_i) + (b-a) \sum_{i=0}^{n} H_i e_i$$

eventual K_i în loc de H_i). Termenul

$$(7.3) \qquad \tau_n = (b-a) \sum_{i=0}^{n} H_i e_i$$

reprezintă eroarea de rotunjire acumulată. Dacă ε este un prag admis pentru numerele e_i, adică $|e_i| \leq \varepsilon$, $i = 0,1,\ldots,n$, atunci

$$(7.4) \qquad |\tau_n| \leq (b-a)\,\varepsilon \sum_{i=0}^{n} |H_i|$$

Dacă, în particular, H_i sînt pozitivi (K_i sînt pozitivi), aşa cum se întîmplă la formulele Newton-Côtes închise pe un număr mic de noduri şi la cîteva formule de tip deschis, obţinem

$$|\tau_n| \leq (b-a)\,\varepsilon \sum_{i=0}^{n} H_i$$

sau, în virtutea unei proprietăţi a coeficienţilor Côtes,

$$(7.5) \qquad |\tau_n| \leq \varepsilon(b-a) \quad .$$

In acest caz, numărul $\varepsilon(b-a)$ nu depinde de h şi, deci, metoda este stabilă.

x

x x

Pe spaţiul Banach $C[a,b]$ considerăm funcţionala $I:C[a,b] \to R$ definită prin

$$(7.6) \qquad I(f) = \int_a^b f(x)dx$$

(faptul că ponderea a fost luată 1 nu este esenţială), care este o funcţională liniară şi continuă.

Fie $(I_n(f))_n$ şirul de aproximaţii ale lui $I(f)$ definit prin

$$(7.7) \qquad I_n(f) = \sum_{i=0}^{n} A_i^{(n)} f(x_i^{(n)}) \quad ,$$

unde nodurile $x_i^{(n)} \in [a,b]$, $i = 0,1,\ldots,n$ (variabila n a şirului indică numărul de noduri). Se pune problema dacă

$$(7.8) \qquad \lim_{n\to\infty} I_n(f) = I(f)$$

pentru orice $f \in C[a,b]$.

Teorema 7.1. Dacă sînt satisfăcute condiţiile: 1° egalitatea (7.8) are loc pentru orice polinom; 2° există $M > 0$ astfel încît

$$(7.9) \qquad \sum_{i=0}^{n} |A_i^{(n)}| \leq M \quad , \qquad (\forall) n \in \mathbb{N}$$

atunci relaţia (7.8) are loc pe spaţiul $C[a,b]$.

Demonstraţie. Fie $f \in C[a,b]$ şi $\varepsilon > 0$. Conform teoremei lui Weierstrass se aproximare uniformă [6] există un polinom $P_\varepsilon(x)$ astfel încît

$$|f(x) - P_\varepsilon(x)| < \varepsilon \quad , \qquad (\forall) x \in [a,b] \ .$$

Dar, conform ipotezei 1°, există n_ε aşa ca

$$|I_n(P_\varepsilon) - I(P_\varepsilon)| < \varepsilon \qquad , \qquad (\forall) \ n > n_\varepsilon \ .$$

Ţinînd seama de (7.9) şi faptul că funcţionalele I şi I_n sînt liniare şi continue, rezultă că

$$|I_n(f)-I(f)| \leq |I_n(f-P_\varepsilon)| + |I_n(P_\varepsilon)-I(P_\varepsilon)| + |I(f-P_\varepsilon)| < \varepsilon M + \varepsilon + \varepsilon(b-a) =$$
$$= [M+1+(b-a)]\varepsilon \ ,$$

oricare ar fi $n > n_\varepsilon$. Ca urmare are loc (7.8) pentru orice $f \in C[a,b]$.

Corolar. Dacă sînt îndeplinite condiţiile :

1° $A_i^{(n)} > 0$, $i = 0,1,\ldots,n$ şi $(\forall) n \in \mathbb{N}$;

2° $\lim_{n\to\infty} m_n = \infty$, unde m_n este gradul de exactitate al formulei de cuadratură $I(f) = I_n(f) + E_n(f)$, atunci (7.8) are loc pe spaţiul $C[a,b]$.

Demonstraţie. Fie $P(x)$ un polinom arbitrar şi k gradul său. Conform ipotezelor există un rang n_k aşa ca $n > n_k$ să implice $m_n > k$. Ca urmare

$$I(P) = I_n(P), \qquad (\forall)\ n > n_k$$

şi deci

$$\lim_{n\to\infty} I_n(P) = I(P)$$

şi aceasta pentru un polinom oarecare. Pe de altă parte, ţinînd seama de 1° şi 2° avem:

$$\sum_{i=0}^{n} |A_i^{(n)}| = \sum_{i=0}^{n} A_i^{(n)} = I_n(1) = I(1) = b-a$$

Aşadar, se poate aplica Teorema 7.1 şi demonstraţia se încheie.

Atît teorema precedentă, cît şi corolarul ei dau condiţii suficiente pentru ca procesul de cuadratură să fie convergent pe spaţiul $C[a,b]$. Formulele de cuadratură ale lui Gauss satisfac ipotezele corolarului.

In practică nu este indicată utilizarea formulelor de cuadratură cu un număr relativ mare de noduri. Am văzut că este recomandabilă utilizarea formulelor iterate. Rezultatul ce urmează se referă la convergenţa unui proces de cuadratură de acest fel.

Teorema 7.2. Fie $a_k = a + k\frac{b-a}{N}$, $k = 0,1,\dots,N$ punctele unei diviziuni echidistante a intervalului $[a,b]$ şi pe fiecare subinterval $[a_k, a_{k+1}]$ o formulă de cuadratură cu gradul de exactitate m de forma $\sum_{i=0}^{n} A_{ki}^{(n)} f(x_{ki}^{(n)})$, unde $x_{ki}^{(n)} \in [a_k, a_{k+1}]$ $(k=0,1,\dots,N-1;\ i = 0,1,\dots,n)$ şi n este un număr natural fixat.

Dacă $A_{ki}^{(n)} > 0$ pentru $k = 0,1,\dots,N$ şi $i = 0,1,\dots,n)$, atunci pentru procesul de cuadratură dat de

$$I_N(f) = \sum_{k=0}^{N-1} \sum_{i=0}^{n} A_{ki}^{(n)} f(x_{ki}^{(n)}) \tag{7.10}$$

are loc convergenţa $\lim_{N\to\infty} I_N(f) = I(f)$ pentru orice $f \in C[a,b]$.

Demonstrație. Fie o funcție arbitrară $f\in C[a,b]$ și $\varepsilon>0$. Deoarece f este uniform continuă pe $[a,b]$, există $\delta_\varepsilon>0$ așa ca $(\forall)x,\ x'\in[a,b]$, $|x-x'|<\delta_\varepsilon$ să avem $|f(x)-f(x')|<\varepsilon$. Fie $h=\frac{b-a}{N}$ lungimea uneia dintre intervalele $[a_k,\ a_{k+1}]$. Evident, există un rang n_ε astfel încît $N>n_\varepsilon$ implică $h<\delta_\varepsilon$. Ca urmare pentru orice punct $\xi_k\in[a_k,a_{k+1}]$ putem scrie:

$$(7.11)\qquad |f(x)-f(\xi_k)|<\varepsilon \quad , \quad (\forall)\ x\in[a_k,a_{k+1}] .$$

Atunci

$$|I_N(f)-I(f)|=\Big|\sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}}f(x)dx-\sum_{k=0}^{N-1}\sum_{i=0}^{n}A_{ki}^{(n)}f(x_{ki}^{(n)})\Big|$$

$$\leq \sum_{k=0}^{N-1}\Big|\int_{a_k}^{a_{k+1}}f(x)dx-\sum_{i=0}^{n}A_{ki}^{(n)}f(x_{ki}^{(n)})\Big|$$

$$\sum_{k=0}^{N-1}\Big(\int_{a_k}^{a_{k+1}}|f(x)-f(\xi_k)|dx+\Big|\int_{a_k}^{a_{k+1}}f(\xi_k)dx-\sum_{i=0}^{n}A_{ki}^{(n)}f(x_{ki}^{(n)})\Big|\Big) .$$

Avînd în vedere (7.11) și faptul că $a_{k+1}-a_k=\sum_{i=0}^{n}A_{ki}^{(n)}$ (egalitate obținută scriind că formula de cuadratură este exactă pentru $f=1$ pe $[a_k,a_{k+1}]$), obținem:

$$|I_N(f)-I(f)|<\varepsilon(b-a)+\sum_{k=0}^{N-1}\Big|\sum_{i=0}^{n}A_{ki}^{(n)}(f(\xi_k)-f(x_{ki}^{(n)}))\Big|\leq$$

$$\leq\varepsilon(b-a)+\sum_{k=0}^{N-1}\sum_{i=0}^{n}A_{ki}^{(n)}|f(\xi_k)-f(x_{ki}^{(n)})|=\varepsilon(b-a)+\varepsilon\sum_{k=0}^{N-1}(a_{k+1}-a_k)=$$

$$=2\varepsilon(b-a),$$

de îndată ce $N>n_\varepsilon$, ceea ce trebuia demonstrat.

Formulele de cuadratură: a dreptunghiurilor, a trapezelor, a lui Simpson (iterată) verifică ipotezele teoremei precedende și, ca urmare, pentru aceste formule are loc convergența lor în sensul specificat mai sus.

Observaţie. Fie $C^*[a,b]$ mulţimea funcţionalelor liniare şi continue pe spaţiul $C[a,b]$, adică dualul acestuia. Evident, $I, I_n, I_N \in C^*[a,b]$. Stim că dualul are o structură de spaţiu Banach faţă de operaţiile liniare naturale şi de norma $\|I\| = \sup_{\|f\| \le 1} |I(f)|$ (a se vedea (I.5.3). Ca urmare se poate vorbi de convergenţa unui şir $(I_n) \subset C^*[a,b]$ în sensul acestei norme, numită şi convergenţa uniformă sau tare a şirului $(I_n)_n$. Dacă termenii I_n, $n \in \mathbb{N}$, şi limita I au semnificaţiile anterioare, se poate arăta [6 ,p.278] că egalitatea

$$\lim_n I_n = I$$

este imposibilă în spaţiul $C^*[a,b]$, adică procesul de cuadratură nu converge uniform.

Tipul de convergenţă prezent în teoremele precedente este un altul: convergenţa punctuală sau slabă. El constă în: $(I_n)_n$ converge punctual la I dacă şirul numeric $(I_n(f))_n$ converge la $I(f)$, oricare ar fi $f \in C[a,b]$. Aşadar, procesul de cuadratură este, în ipotezele specificate, punctual convergent.

§ 8. Exerciţii

1. Să se utilizeze formulele dreptunghiului, trapezului şi Simpson (simplă) pentru a aproxima integralele:

a) $\int_0^1 e^{-x}\,dx$,

b) $\int_1^2 \ln x\,dx$,

c) $\int_0^{\frac{\pi}{2}} \cos^2 x\,dx$.

Să se găsească o limită pentru eroarea absolută în fiecare caz. Să se compare cu valorile exacte ale acestor integrale.

2. Să se utilizeze formulele dreptunghiurilor, trapezelor şi Simpson (iterată) pentru a aproxima cu o eroare mai mică ca 10^{-3} următoarele integrale:

a) $\int_0^1 \frac{dx}{1+x}$,

b) $\int_0^{2\pi} \frac{dx}{2-\cos x}$,

c) $\int_1^2 e^x \cos x \, dx$;

Pentru fiecare integrală se vor compara valorile aproximative între ele şi cu valoarea exactă.

3. Cu aceleaşi metode şi cu aceeaşi eroare ca în exerciţiul precedent, să se aproximeze integralele:

a) $\int_0^{0,5} \frac{dx}{1+x^7}$

b) $\int_0^{\frac{\pi}{2}} \sin x^2 \, dx$, $\int_0^{\frac{\pi}{2}} \cos x^2 \, dx$,

c) $\int_0^{\frac{\pi}{4}} \sin(\sin x) dx$, $\int_0^{\frac{\pi}{4}} \sin(\cos x) dx$, etc.,

d) $\int_0^{\frac{\pi}{2}} e^{\cos x} \, dx$, $\int_0^{\frac{\pi}{2}} e^{\sin x} \, dx$,

e) $\int_0^{\frac{\pi}{2}} x^5 \sin \frac{1}{x} \, dx$,

f) $\int_0^1 \cos \frac{x}{1+x} \, dx$,

g) $\int_0^{\frac{\pi}{2}} \sqrt{1-0,25 \sin^2 x} \, dx$.

4. O rachetă este lansată de la o bază. Acceleraţia sa este înregistrată în primele 80 secunde şi este dată de tabelul

t(sec)	0	10	20	30	40	50	60	70	80
a(m/sec^2)	30,00	31,63	33,44	35,75	37,75	40,33	43,29	46,69	50,67

Să se găsească viteza şi înălţimea rachetei la momentul t = 80.

5. Ecuaţia care modelează funcţionarea unui circuit electric compus dintr-o sursă de curent, o rezistenţă şi un condensator este

$$E = R\,i + \frac{1}{C}\int_0^t i \, dt .$$

Presupunînd că pentru un circuit dat avem $R = 0,1\ \Omega$, $C = 1$ F şi $i = \ln(2t+1)$ A , să se găsească E după 2 secunde, utilizînd formula lui Simpson iterată cu pasul h = 0,1.

6. Să se calculeze lungimea elipsei

$$\frac{x^2}{4} + \frac{y^2}{9} = 1$$

cu cinci zecimale exacte.

7. Numărul numerelor prime din intervalul (a,b) este aproximat de

$$\int_a^b \frac{dx}{\ln x}$$

Să se utilizeze acest rezultat pentru a = 100 şi b = 200 şi să se compare cu valoarea exactă.

8. Să se determine coeficienţii formulei de cuadratură

$$\int_0^1 f(x)dx = A_o f(o) + A_1 f(1) + B_o f'(o) + B_1 f'(1) + E(f)$$

astfel încît formula să aibă gradul de exactitate cel puţin egal cu trei. Cît este gradul de exactitate al formulei obţinuta ?

9. Să se deducă o formulă de cuadratură cu două noduri de tip Gauss pentru integralele de forma:

a) $\int_0^1 \sqrt{x}\, f(x)dx$,

b) $\int_0^{\frac{\pi}{2}} \sin x\, f(x)dx$,

c) $\int_0^1 e^x f(x)dx$.

10. Să se aproximeze cu o eroare mai mică ca 10^{-4} următoarele integrale:

a) $\int_{-\infty}^{+\infty} \frac{e^{-x^2}}{1+x^2} dx$,

b) $\int_{-\infty}^{+\infty} e^{-x^2} \sin x\, dx$, $\int_{-\infty}^{+\infty} e^{-x^2} \cos x\, dx$,

c) $\int_{-\infty}^{+\infty} e^{-x^2} \operatorname{arctg} x\, dx$,

utilizînd formula de cuadratură Gauss-Hermite.

11. Să se aproximeze cu o eroare mai mică ca 10^{-2} integralele:

a) $\int_0^{\infty} e^{-x} \sin x \, dx$,

b) $\int_0^{\infty} e^{-x} \ln(1+x) dx$,

utilizînd formula de cuadratură Gauss-Laguerre.

12. Să se completeze tabelul de valori ale funcţiei eulierene $\Gamma(x)$

x	1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2
$\Gamma(x)$	1				$\frac{\sqrt{\pi}}{2} \approx 0,8862$					1

utilizînd formula

$$\int_0^{\infty} e^{-x} f(x) dx \approx \frac{2+\sqrt{2}}{4} f(2-\sqrt{2}) + \frac{2-\sqrt{2}}{4} f(2+\sqrt{2})$$

(formula Gauss-Laguerre cu două noduri).

13. Fie formula de cuadratură Gauss-Cebîşev cu 3 noduri:

$$\int_{-1}^{+1} \frac{f(x)dx}{\sqrt{1-x^2}} = \frac{\pi}{3}\left[f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right)\right] + E(f) .$$

a) Să se verifice că formula are gradul de exactitate 5.

b) Să se aproximeze integrala pentru $f(x) = \ln \frac{2-x}{2+x}$.

14. Utilizînd polinoamele lui Cebîşev de prima specie, să se stabilească următoarea formulă de cuadratură de tip Gauss:

$$\int_{-1}^{+1} \frac{f(x)}{\sqrt{1-x^2}} dx = \frac{\pi}{n} \sum_{k=1}^{n} f\left(\cos \frac{2k-1}{2n} \pi\right) + \frac{2\pi}{2^{2n}(2n)!} f^{(2n)}(\xi) ,$$

unde $\xi \in (-1,1)$ (Formula de cuadratură Gauss-Cebîşev).

15. Fie formula de cuadratură

$$\int_{-1}^{+1} f(x)dx = C_1 f(x_1) + C_2 f(x_2) + C_3 f(x_3) + E(f) .$$

Să se determine coeficienţii C_1, C_2 şi C_3 precum şi nodurile $x_1, x_2, x_3 \in [-1,1]$ astfel încît să fie satisfăcute condiţiile:

1°. $C_1 = C_2 = C_3$,

2° formula este exactă pentru polinoamele de grad mai mic sau egal cu 3 (Formula lui Cebîşev cu 3 noduri).

16. Să se alcătuiască algoritmul şi schema logică pentru

a) formula dreptunghiurilor,

b) formula trapezelor,

c) formula de cuadratură Gauss-Legendre.

Capitolul VII

DERIVARE NUMERICA

Calculul derivatelor unei funcţii într-un punct dat apare frecvent în aplicaţii. Aproximarea acestor derivate devine, ca urmare, o problemă de importanţă majoră.

O formulă de derivare numerică se poate obţine - ca şi în cazul formulelor de cuadratură - prin derivarea unei formule de interpolare o singură dată sau de mai multe ori. Menţionăm însă că acesta nu este singurul mod de a proceda. Si în cazul derivării numerice vom folosi polinoamele ca funcţii de interpolare.Această preferinţă se explică prin faptul că mulţimea polinoamelor este densă în spaţiul $C[a,b]$, adică pentru orice funcţie continuă definită pe $[a,b]$ există un polinom ce o aproximează uniform oricît de bine pe acest interval. Un alt motiv constă în faptul că derivata unui polinom este tot un polinom şi poate fi uşor calculată.

§ 1. Formule de derivare numerică în noduri arbitrare

Fie date nodurile $x_0, x_1, \ldots, x_n \in [a,b]$ şi valorile unei funcţii f în aceste puncte. Formula de interpolare a lui Lagrange pe aceste noduri este

(1.1) $$f(x) = L(x_0,\ldots,x_n;f|x) + f[x_0,x_1,\ldots,x_n,x] \cdot \omega(x) \quad ,$$

unde

(1.2) $$\omega(x) = (x-x_0)(x-x_1),\ldots,(x-x_n)$$

şi

$$(1.3)\qquad L(x_o,x_1,\dots,x_n;\ f|x) = \sum_{i=o}^{n} L_i(x)\ f(x_i)$$

cu

$$(1.4)\qquad L_i(x) = \frac{\omega(x)}{(x-x_i)\ \omega'(x_i)}\ ,\qquad i = o,1,\dots,n\ .$$

Ca urmare, cunoscînd valorile funcţiei f în noduri, formula aproximativă

$$(1.5)\qquad f^{(m)}(x) \approx L^{(m)}(x_o,x_1,\dots,x_n;\ f|x),\qquad o \le m \le n\ ,$$

sau echivalent,

$$(1.6)\qquad f^{(m)}(x) \approx \sum_{i=o}^{n} L_i^{(m)}(x).f(x_i),\qquad o \le m \le n\ ,$$

reduce calculul derivatelor funcţiei date de ordin pînă la n la calculul mai simplu al derivatelor polinomului Lagrange. Formulele (1.5) sau (1.6) pot fi utilizate formal pentru orice funcţie dată printr-un tabel de valori în noduri. In acest caz, însă, nu se poate evalua eroarea acestor egalităţi aproximative. Mai mult, rezultatele oferite de (1.5) sau (1.6) pot fi eronate, căci prin punctele $(x_i,\ f(x_i))$, $i = o,1,\dots,n$, pot trece curbe a căror alură să fie cu totul diferită de cea a polinomului Lagrange. Va trebui să impunem condiţii suplimentare relativ la funcţia f pentru a obţine expresii sau evaluări ale erorii formulelor (1.5) sau (1.6).

Teorema 1.1. Fie $f:[a,b]\to R$, nodurile $x_o,x_1,\dots,x_n \in [a,b]$ şi $m \in N$, $m \le n$. Dacă $f \in C^{n+m+1}[a,b]$, atunci există punctele ξ_k, $k = o,1,\dots,m$, aparţinînd intervalului minimal ce conţine $x,x_o,\dots,x_n$ astfel ca

$$(1.7)\qquad f^{(m)}(x) = \sum_{i=o}^{n} L_i^{(m)}(x).f(x_i) + \sum_{k=o}^{m} \frac{k!}{(n+k+1)}\ C_m^k\ f^{(n+k+1)}(\xi_k)\,\omega^{(m-k)}(x)\ .$$

Demonstrație. Să derivăm de m ori formula (1.1) de interpolare a lui Lagrange:

$$f^{(m)}(x)=\sum_{i=o}^{n} L_i^{(m)}(x).f(x_i)+\sum_{k=o}^{m} C_m^k \frac{d^k}{dx^k} f[x_o,\ldots,x_n,x]\ \omega^{(m-k)}(x).$$

Dar, în conformitate cu (V.3.17), în ipotezele noastre avem:

$$\frac{d^k}{dx^k} f[x_o,\ldots,x_n,x]=\frac{k\,!}{(n+k+1)!} f^{(n+k+1)}(\xi_k) \quad ,$$

cu $\xi_k \in (\min\{x,x_o,x_1,\ldots,x_n\},\ \max\{x,x_o,x_1,\ldots,x_n\})$ și $k=o,1,\ldots,m$. Formula (1.7) se obține combinînd ultimele două egalități.

Observație. Pentru derivata de ordin întîi a funcției $f\in C^{n+2}[a,b]$ se obține, luînd $m=1$ în (1.7),

$$(1.8)\qquad f'(x)=\sum_{i=o}^{n} L_i'(x).f(x_i)+ \frac{1}{(n+1)!} f^{(n+1)}(\xi)\ \omega'(x)+\frac{1}{(n+2)!} f^{(n+2)}(\eta)\ \omega(x)\ .$$

Această formulă ia o formă mai simplă atunci cînd x este unul dintre noduri, $x=x_j$:

$$(1.9)\qquad f'(x_j)=\sum_{i=o}^{n} L_i'(x_j)\ f(x_i)+\frac{1}{(n+1)!} f^{(n+1)}(\xi_j)\ \omega'(x_j)\ .$$

Atunci cînd derivatele funcției f se calculează într-un punct x situat în afara intervalului nodurilor, restul formulei (1.7) poate fi pus sub o formă mai simplă.

Teorema 1.2. Fie $f\in C^{n+1}[a,b]$ și nodurile $x_o,x_1,\ldots,x_n\in[a,b]$. Dacă $x\in[a,b]$ și $x\notin(\min x_i,\ \max x_i)$, atunci pentru orice $m\le n$ există ξ_m aparținînd intervalului punctelor $x,x_o,\ldots,x_n$ astfel ca

$$(1.10)\qquad f^{(m)}(x)=\sum_{i=o}^{n} L_i^{(m)}(x).f(x_i)+\frac{1}{(n+1)!} f^{(n+1)}(\xi_m)\ \omega^{(m)}(x).$$

<u>Demonstrație</u>. In conformitate cu teorema lui Rolle, rădăcinile derivatelor polinomului $\omega(x)$ se găsesc în intervalele determinate de rădăcinile lui $\omega(x)$. Ca urmare, din faptul că x nu aparține intervalului nodurilor rezultă că $\omega^{(m)}(x) \neq 0$, $m \leq n$. Există atunci un număr $C_m \in R$ astfel ca

$$(1.11) \qquad f^{(m)}(x) - \sum_{i=0}^{n} L_i^{(m)}(x).f(x_i) - C_m\, \omega^{(m)}(x) = 0$$

Să considerăm funcția $\varphi : [a,b] \longrightarrow R$ definită de

$$\varphi(t) = f(t) - \sum_{i=0}^{n} L_i(t).f(x_i) - C_m\, \omega(t) \quad .$$

Deoarece $\varphi \in C^{n+1}[a,b]$ și se anulează în $x_0, x_1, \ldots, x_n$ rezultă, conform teoremei lui Rolle, că $\varphi^{(m)}$ se anulează de $(n+1) - m$ ori în intervalul acestor noduri. Din modul cum a fost introdus numărul C_m avem că $\varphi^{(m)}(x) = 0$ și, deci, $\varphi^{(m)}$ se anulează în $(n+1) - m+1$ puncte. Utilizînd din nou teorema lui Rolle va rezulta că derivata de ordin $(n+1) - m$ a lui $\varphi^{(m)}$, adică $\varphi^{(n+1)}$, se anulează într-un punct ξ_m ce aparține intervalului punctelor $x, x_0, \ldots, x_n$. Așadar

$$\varphi^{(n+1)}(\xi_m) = 0$$

sau, ținînd seama de definiția funcției φ,

$$f^{(n+1)}(\xi_m) - (n+1)!\, C_m = 0 \quad .$$

Deci

$$C_m = \frac{1}{(n+1)!} f^{(n+1)}(\xi_m)$$

și înlocuind în (1.11) obținem relația (1.10). Q.e.d.

x

x x

Dacă în locul formulei (1.1) se ia ca punct de plecare formula de interpolare a lui Newton

$$(1.1') \quad f(x)=f[x_0]+\sum_{i=1}^{n}(x-x_0)(x-x_1),\dots,(x-x_{i-1})f[x_0,x_1,\dots,x_i]+$$
$$+\omega(x).f[x_0,\dots,x_n,x] \quad ,$$

considerațiile de mai sus se pot reface pas cu pas. Formula aproximativă (1.6) se înlocuieşte cu:

$$(1.6') \quad f^{(m)}(x)\approx\sum_{i=1}^{n}\left[(x-x_0),\dots,(x-x_{i-1})\right]^{(m)}.f[x_0,x_1,\dots x_i],$$
$$1\le m\le n\,.$$

Observăm că, în timp ce (1.6) exprimă derivata cu ajutorul valorilor funcției f în noduri, formula (1.6') exprimă aceeaşi derivată cu ajutorul diferențelor divizate ale lui f pe nodurile $x_0,x_1,\dots,x_n$. In privința restului nu apar nici un fel de modificări în formulele de mai sus, căci este aceeaşi în cele două formule de interpolare luate ca punct de plecare.

§ 2. Formula de derivare numerică în noduri echidistante

Să presupunem că nodurile $x_0,x_1,\dots,x_n$ sînt echidistante, situație frecventă în aplicații. Fie h pasul rețelei acestor noduri, aşadar $x_{i+1}-x_i=h$ $(i=0,1,\dots,n-1)$, ceea ce este echivalent cu

$$(2.1) \qquad x_i=x_0+ih \quad , \qquad i=0,1,\dots,n.$$

Punînd

$$(2.2) \qquad x=x_0+th$$

şi ținînd seama de (1.2), obținem:

$$(2.3) \qquad \omega(x)=h^{n+1}\,\pi_n(t)\ ,$$

unde $\pi_n(t) = t(t-1)\dots(t-n)$ este polinomul factorial. De asemenea, avem:

$$\omega'(x_i) = (x_i - x_0)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n),$$

adică, datorită relaţiilor (2.1):

$$\omega'(x_i) = (-1)^{n-i}\, h^n\, i!\,(n-i)! \quad . \tag{2.4}$$

Ca urmare, în virtutea relaţiilor (1.3) şi (1.4),

$$L(x_0,x_1,\dots,x_n;\, f|x) = \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!(n-i)!} \cdot \frac{\pi_n(t)}{t-i}\, f(x_i) \quad .$$

Formula (1.6) devine

$$f^{(m)}(x) \approx \frac{1}{h^m} \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!(n-i)!} \cdot \frac{d^m}{dt^m}\left[\frac{\pi_n(t)}{t-i}\right] f(x_i), \quad 0 \le m \le n . \tag{2.5}$$

Pe de altă parte, avînd în vedere formula de interpolare a lui Newton progresivă (V.4.7), avem:

$$f^{(m)}(x) \approx \frac{1}{h^m} \sum_{i=1}^{n} \frac{1}{i!} \frac{d^m}{dt^m}\left[\pi_{i-1}(t)\right] \cdot \Delta^i f(x_0), \quad 1 \le m \le n . \tag{2.6}$$

In privinţa restului - care este acelaşi, indiferent că este vorba de (2.5) sau (2.6) - vom acorda atenţie doar cazului $m=1$, adică formulei de derivare numerică pentru derivata de ordinul întîi. Mai mult, vom considera că x este unul dintre noduri. In conformitate cu (1.9), (2.4) şi (2.5), obţinem:

$$f'(x_j) = \frac{1}{h} \sum_{i=0}^{n} \frac{(-1)^{n-i}}{i!(n-i)!} \frac{d}{dt}\left[\frac{\pi_n(t)}{t-i}\right] . f(x_i) + \tag{2.7}$$

$$+ (-1)^{n-j} h^n \frac{j!(n-j)!}{(n+1)!} f^{(n+1)}(\xi_j), \quad j=0,1,\dots,n .$$

Să punem în evidenţă cîteva cazuri particulare ale formulei (2.7). In prealabil să convenim asupra notaţiilor: $y_i = f(x_i)$, $y'_i = f'(x_i)$ etc. $(i=0,1,\dots,n)$.

<u>$n = 1$</u> :

$$\begin{aligned} y'_0 &= \frac{1}{h}(y_1 - y_0) - \frac{h}{2} f''(\xi) , \quad x_0 < \xi < x_1 , \\ y'_1 &= \frac{1}{h}(y_1 - y_0) + \frac{h}{2} f''(\xi) , \quad x_1 < \xi < x_1 . \end{aligned} \tag{2.8}$$

<u>n = 2</u> :

$$
(2.9)\qquad
\begin{aligned}
y_0' &= \frac{1}{2h}(-3y_0 + 4y_1 - y_2) + \frac{h^2}{3} f'''(\xi),\\
y_1' &= \frac{1}{2h}(y_2 - y_0) - \frac{h^2}{6} f'''(\xi) ,\\
y_2' &= \frac{1}{2h}(y_0 - 4y_1 + 3y_2) + \frac{h^2}{3} f'''(\xi) .
\end{aligned}
$$

<u>n = 3</u> :

$$
(2.10)\qquad
\begin{aligned}
y_0' &= \frac{1}{6h}(-11 y_0 + 18y_1 - 9y_2 + 2y_3) - \frac{h^3}{4} f^{(4)}(\xi) ,\\
y_1' &= \frac{1}{6h}(-2y_0 - 3y_1 + 6y_2 - y_3) + \frac{h^3}{12} f^{(4)}(\xi) ,\\
y_2' &= \frac{1}{6h}(y_0 - 6y_1 + 3y_2 + 2y_3) - \frac{h^3}{12} f^{(4)}(\xi) ,\\
y_3' &= \frac{1}{6h}(-2y_0 + 9y_1 - 18y_2 + 11y_3) + \frac{h^3}{4} f^{(4)}(\xi).
\end{aligned}
$$

Calculul numeric al derivatei secunde este o operaţie mai puţin sigură decît cel al primei derivate. Obţinerea derivatei secunde în noduri se poate face prin punerea la punct a unei formule de tipul (2.7) relativ la această derivată. Vom indica însă şi un alt mod de a proceda, ce putea fi folosit şi pentru derivata de ordinul întîi; în principiu, va fi utilizată formula lui Taylor.

<u>n = 2</u> :

$$
(2.11)\qquad
\begin{aligned}
y_0'' &= \frac{1}{h^2}(y_0 - 2y_1 + y_2) - h\, f'''(\xi),\\
y_1'' &= \frac{1}{h^2}(y_0 - 2y_1 + y_2) - \frac{h^2}{12} f^{(4)}(\xi) ,\\
y_2'' &= \frac{1}{h^2}(y_0 - 2y_1 + y_2) + h\, f'''(\xi) .
\end{aligned}
$$

Prima şi ultima dintre aceste formule rezultă din (1.10). Rămîne să demonstrăm a doua din grupul de formule (2.11). Utilizăm formula Taylor de ordinul trei în punctul x_1, cu restul sub

forma lui Lagrange şi calculăm $f(x_0)$ şi $f(x_2)$ ţinînd seama că $x_0 = x_1 - h$ şi $x_2 = x_1 + h$; avem:

$$y_0 = y_1 - h\,y_1' + \frac{h^2}{2}\,y_1'' - \frac{h^3}{6}\,y_1''' + \frac{h^4}{24}\,f^{(4)}(\xi') ,$$

$$y_2 = y_1 + h\,y_1' + \frac{h^2}{2}\,y_1'' + \frac{h^3}{6}\,y_1''' + \frac{h^4}{24}\,f^{(4)}(\xi'')$$

Prin adunare, se ajunge la

$$y_0 + y_2 = 2y_1 + h^2 y_1'' + \frac{h^4}{12}\cdot\frac{1}{2}\left[f^{(4)}(\xi') + f^{(4)}(\xi'')\right]$$

şi deoarece există ξ, $x_0 < \xi < x_2$, astfel ca $f^{(4)}(\xi) = \frac{1}{2}(f^{(4)}(\xi') + f^{(4)}(\xi''))$, se obţine

$$y_0 + y_2 = 2y_1 + h^2 y_1'' + \frac{h^4}{12}\,f^{(4)}(\xi)$$

de unde

$$y_1'' = \frac{1}{h^2}\,(y_0 - 2y_1 + y_2) - \frac{h^2}{12}\,f^{(4)}(\xi) , \quad x_0 < \xi < x_2 .$$

In mod similar, pentru cazul

<u>n = 3</u> :

(2,12)
$$\begin{aligned}
y_0'' &= \frac{1}{h^2}\,(2y_0 - 5y_1 + 4y_2 - y_3) + \frac{11}{12}\,h^2 f^{(4)}(\xi) ,\\
y_1'' &= \frac{1}{h^2}\,(y_0 - 2y_1 + y_2) - \frac{h^2}{12}\,f^{(4)}(\xi) ,\\
y_2'' &= \frac{1}{h^2}\,(y_1 - 2y_2 + y_3) - \frac{h^2}{12}\,f^{(4)}(\xi) ,\\
y_3'' &= \frac{1}{h^3}\,(-y_0 + 4y_1 - 5y_2 + 2y_3) + \frac{11}{12}\,h^2\,f^{(4)}(\xi) .
\end{aligned}$$

<u>Observaţie</u>. Cu creşterea lui n - şi , corespunzător, a clasei funcţiei f - creşte şi ordinul resturilor formulelor relativ la h. Dacă, însă, creşte ordinul m al derivatei, ordinul resturilor relativ la h descreşte. Expresiile derivatelor în nodurile ce se află mai aproape de mijlocul intervalului (x_0, x_n) sînt mai simple decît în extremităţile lui. Pentru n par, restul deri-

vatei de ordin par în nodul din mijloc este de ordin mai mare cu o unitate. De aceea se recomandă, pe cît este posibil, să se folosească formulele de derivare numerică pe noduri situate simetric față de punctul în care se caută derivata. Cele mai simple şi mai utilizate formule de derivare numerică sînt (2.8) şi formulele din mijlocul grupurilor (2.9) şi (2.11).

<u>Exemplul 2.1</u>. Să se calculeze aproximativ $f'(1{,}06)$ şi $f'(1{,}06)$ utilizînd valorile funcției f date în tabelul ce urmează:

x	1,02	1,04	1,06	1,08	1,10
$f(x)$	5,85440	5,99408	6,13738	6,28439	6,43516

Utilizînd formulele (2.8) avem:

$$f'(1{,}06) \approx \frac{1}{h}\left[f(1{,}08)-f(1{,}06)\right]=\frac{1}{0{,}02}(6{,}28439-6{,}13738)=7{,}3505$$

şi

$$f'(1{,}06) \approx \frac{1}{h}\left[f(1{,}06)-f(1{,}04)\right]=\frac{1}{0{,}02}(6{,}13738-5{,}99408)=7{,}165.$$

Dacă folosim a doua formulă (2.9), obținem

$$f'(1{,}06) \approx \frac{1}{2h}\left[f(1{,}08)-f(1{,}04)\right]=\frac{1}{0{,}04}(6{,}28439-5{,}99408)=7{,}25775.$$

Funcția pe care am avut-o în vedere în mod tacit este $f(x) = e^x + x^3 + x+1$. Comparînd valorile găsite pentru $f'(1{,}06)$ cu valoarea exactă egală cu 7,25717, observăm că a treia valoare calculată este mai bună comparativ cu primele două. Aceasta este în concordanță cu faptul că restul ultimei formule de derivare numerică utilizată mai sus este de ordinul lui h^2, pe cînd resturile primelor două sînt de ordinul lui h.

In privința aproximării numărului $f''(1{,}06)$ vom utiliza a doua formulă (2.11). Avem

$$f''(1{,}06) \approx \frac{1}{h^2}\left[f(1{,}04)-2f(1{,}06)+f(1{,}08)\right]=$$

$$=\frac{1}{0{,}0004}(5{,}99408-2.6{,}13738+6{,}28439) = 9{,}275 .$$

Valoarea exactă este $f'(1,06) = 9,24637$.

Formulele de derivare numerică prezentate mai sus necesită cunoașterea valorilor funcției în noduri. Uneori este preferabilă o formulă care face apel la diferențele finite ale funcției f relativ la rețeaua de noduri echidistante. O astfel de formulă se poate obține prin derivarea - de numărul dorit de ori - a oricărei dintre formulele de interpolare pe noduri echidistante stabilite în Cap.V. Astfel, formula

$$(2.13)\qquad f'(x_0) = \frac{1}{h}\left(\Delta f(x_0) - \frac{1}{2}\Delta^2 f(x_0) + \frac{1}{3}\Delta^3 f(x_0) + \dots + (-1)^{n-1}\frac{1}{n}\Delta^n f(x_0)\right) + (-1)^n \frac{h^4}{n+1} f^{(n+1)}(\xi)$$

se obține luînd $m = 1$ și $x = x_0$ în (2.6) și utilizînd pentru rest expresia din (2.7). Observăm că, dacă facem succesiv $n = 1$ și $n=2$ în (2.13), obținem prima formulă din (2.8) și prima din (2.9), respectiv. De asemenea, derivata de ordin întîi în nodul x_1 se obține luînd în (2.6) $x=x_1$, etc. In sfîrșit, să mai observăm că prin derivarea unei formule de interpolare cu diferențe centrale vom ajunge la o formulă de derivare numerică ce ține cont atît de valorile funcției situate la stînga punctului x cît și de cele de la dreapta lui.

Dacă derivăm formula de interpolare a lui Gauss progresivă, formula (V . 4.13) și luăm $t = 0$, obținem

$$(2.14)\qquad f'(x_0) = \frac{1}{h}\Big[\delta f(x_{\frac{1}{2}}) - \frac{1}{2!}\delta^2 f(x_0) - \frac{1}{3!}\delta^3 f(x_{\frac{1}{2}}) + \dots + (-1)^{n-1}\frac{[(n-1)!]^2}{(2n-1)!}\delta^{2n-1} f(x_{\frac{1}{2}}) + (-1)^n \frac{(n-1)!\,n!}{(2n)!}\delta^{2n} f(x_0)\Big] + (-1)^n \frac{(n!)^2}{(2n+1)!} h^{2n} f^{(2n+1)}(\xi) .$$

§ 3. Observații asupra influenței erorilor de rotunjire

In formulele de derivare numerică (2.5) şi (2.6) cu pasul constant h, valorile funcţiei f se împart la h^m, unde m este ordinul derivatei calculate. Ca urmare, pentru valori mici ale lui h erorile de rotunjire comise în calculul valorilor lui f în noduri capătă o semnificaţie deosebită.

Să considerăm formulele de derivare numerică

(3.1) $$f'(x_0) = \frac{1}{2h}\left[f(x_0+h) - f(x_0-h)\right] - \frac{h^2}{6} f'''(\xi)$$

şi

(3.2) $$f''(x_0) = \frac{1}{h^2}\left[f(x_0-h) - 2f(x_0) + f(x_0+h)\right] - \frac{h^2}{12} f^{(4)}(\xi),$$

care sînt a doua formulă (2.9) şi, respectiv, a doua formulă (2.11) cu nodurile notate astfel ca să fie pusă în evidenţă „simetria" acestor formule. Fie ε_{-1}, ε_0 şi ε_1 erorile de rotunjire comise atunci cînd sînt calculate valorile funcţiei în nodurile x_0-h, x_0 şi x_0+h; avem

$$f(x_0-h) = \bar{f}(x_0-h) + \varepsilon_{-1},$$
$$f(x_0) = \bar{f}(x_0) + \varepsilon_0,$$
$$f(x_0+h) = \bar{f}(x_0+h) + \varepsilon_1$$

(semnificaţia unor notaţii fiind evidentă). Să presupunem că $|\varepsilon_{-1}|, |\varepsilon_0|, |\varepsilon_1| \le \varepsilon$ şi fie $M_k = \sup_{x\in[x_0-h,x_0+h]} |f^{(k)}(x)|$, cu $k = 3,4$. Atunci avem:

$$f'(x_0) = \frac{1}{2h}(\bar{f}(x_0+h) - \bar{f}(x_0-h)) + \frac{1}{2h}(\varepsilon_1 - \varepsilon_{-1}) - \frac{h^2}{6} f'''(\xi)$$

şi

$$f''(x_0) = \frac{1}{h^2}(\bar{f}(x_0-h) - 2\bar{f}(x_0) + \bar{f}(x_0+h)) + \frac{1}{h^2}(\varepsilon_{-1} - 2\varepsilon_0 + \varepsilon_1) - \frac{h^2}{12} f^{(4)}(\xi).$$

Evident, dacă h este mic termenii datoraţi erorilor de rotunjire pot fi foarte mari. Eroarea totală a formulelor

$$f'(x_0) \approx \frac{1}{2h} (\bar{f}(x_0 + h) - \bar{f}(x_0 - h))$$

şi

$$f''(x_0) \approx \frac{1}{h^2} (\bar{f}(x_0 - h) - 2\bar{f}(x_0) + \bar{f}(x_0 + h))$$

este limitată de

$$\varepsilon'(h) = \frac{\varepsilon}{h} + \frac{h^2}{6} M_3$$

şi, respectiv,

$$\varepsilon''(h) = \frac{4\varepsilon}{h^2} + \frac{h^2}{12} M_4 \quad .$$

Prin calcule elementare se constată că $\varepsilon'(h)$ şi $\varepsilon''(h)$ sînt minime pentru

$$h = h' = \left(\frac{3\varepsilon}{M_3}\right)^{\frac{1}{3}} \quad \text{şi} \quad h = h'' = 2\left(\frac{3\varepsilon}{M_4}\right)^{\frac{1}{4}}$$

şi valorile lor minime sînt, respectiv,

$$\varepsilon' = \frac{3}{2}\left(\frac{\varepsilon^2 M_3}{3}\right)^{\frac{1}{3}} \quad \text{şi} \quad \varepsilon'' = 2\left(\frac{\varepsilon M_4}{3}\right)^{\frac{1}{2}} \quad .$$

In practică nu vom putea calcula un astfel de h optimal în aproximarea derivatei respective, deoarece nu avem nici o cunoştinţă asupra derivatelor de ordin trei sau patru.

O tehnică foarte utilă în mai multe situaţii ce apar în analiza numerică este aşa-numitul <u>procedeu de extrapolare a lui Richardson</u>. Vom ilustra acest procedeu prin obţinerea unei formule de derivare numerică pentru f" cu restul de ordinul lui h^4. Considerăm următoarele dezvoltări Taylor de ordin cinci în jurul punctului x_0:

$$f(x_0+h) = f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + \frac{h^3}{6} f'''(x_0) + \frac{h^4}{24} f^{(4)}(x_0) + \frac{h^5}{120} f^{(5)}(x_0) + \frac{h^6}{720} f^{(6)}(\xi') \quad ,$$

$$f(x_0-h) = f(x_0) - hf'(x_0) + \frac{h^2}{2} f''(x_0) - \frac{h^3}{6} f'''(x_0) + \frac{h^4}{24} f^{(4)}(x_0) - \frac{h^5}{120} f^{(5)}(x_0) + \frac{h^6}{720} f^{(6)}(\xi'') \quad ,$$

cu $x_0 - h < \xi', \xi'' < x_0 + h$. Adunînd şi ţinînd seama că există $\xi_1 \in (x_0 - h, x_0 + h)$ astfel încît $\frac{1}{2}(f^{(6)}(\xi') + f^{(6)}(\xi'')) = f^{(6)}(\xi_1)$, obţinem

$$\frac{1}{h^2}\left[f(x_0+h) - 2f(x_0) + f(x_0-h)\right] = f''(x_0) + \frac{h^2}{12} f^{(4)}(x_0) + \frac{h^4}{360} f^{(6)}(\xi_1).$$

Inlocuim aici h cu qh, numărul q fiind arbitrar, dar diferit de ± 1, avem:

$$\frac{1}{q^2h^2}\left[f(x_0+qh) - 2f(x_0) + f(x_0-qh)\right] = f''(x_0) + \frac{q^2h^2}{12} f^{(4)}(x_0) + \frac{q^4h^4}{360} f^{(6)}(\xi_2).$$

Eliminînd termenul în h^2 din ultimele egalităţi, sîntem conduşi la formula

$$(3.3) \quad f''(x_0) = \frac{1}{h^2q^2(q^2-1)}\left[-f(x_0+qh) + q^4 f(x_0+h) - (2q^4-2)f(x_0) + q^4 f(x_0-h) - f(x_0-qh)\right] + \frac{q^2h^4}{(q^2-1).360}\left[f^{(6)}(\xi_1) - q^2 f^{(6)}(\xi_2)\right]$$

al cărei rest este de ordinul lui h^4.

In încheiere observăm că derivarea numerică ca metodă de aproximare nu-i stabilă întrucît valori mici pentru h conduc la erori mari. In practică această operaţie este evitată atunci cînd este posibil. Totuşi formulele indicate mai sus au aplicaţii în aproximarea soluţiilor ecuaţiilor diferenţiale ordinare şi cu derivate parţiale.

§ 4. Exerciţii

1. Să se aproximeze f'(10,032), f'(10,022) şi f'(10,015) utilizînd tabelul de valori.

x	10,01	10,02	10,03
f(x)	117,2535	117,5071	118,0308

2. Fie $f(x) = 2^x \sin x$. Să se aproximeze f'(1,05) utilizînd $H = 0,05$ şi $h = 0,01$ în formula (3.1) şi tabelul de valori

x	1,00	1,04	1,06	1,10
f(x)	1,6829420	1,7732994	1,8188014	1,9103448

3. Fie funcţia $f(x) = 3x\,e^x - \cos x$. Să se aproximeze $f''(1,3)$ utilizînd formula (3,2), tabelul de valori

x	1,20	1,29	1,30	1,31	1,40
f(x)	11,59006	13,78176	14,04276	14,30741	16,86187

4. Distanţa $s = s(t)$ parcursă de un punct în mişcare rectilinie este dată de tabelul:

t(sec)	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08
s(cm)	0,000	1,519	6,031	13,397	23,396	35,721	50,000	65,798	82,635

Utilizînd diferenţele finite pînă la ordinul cinci inclusiv să se calculeze aproximativ viteza <u>v</u> şi acceleraţia <u>a</u> pe care le are punctul mobil la momentele $t_1 = 0$, $t_2 = 0,01$ şi $t_3 = 0,02$. Să se compare rezultatele găsite cu valorile exacte, legea mişcării fiind dată de $s(t) = 100(1-\cos\frac{50\pi t}{9})$.

5. Ecuaţia care modelează funcţionarea unui circuit electric în care se leagă în serie o rezistenţă şi o bobină este

$$E = L\frac{di}{dt} + Ri .$$

Presupunînd că $R = 0,142\ \Omega$, $L = 0,98$ H şi intensitatea curentului este dată de

t(sec)	1,00	1,01	1,02	1,03	1,04
i(A)	3,10	3,12	3,14	3,18	3,24

să se calculeze E la momentele t=1; 1,01; 1,02; 1,03 şi 1,04.

6. Să se deducă o formulă de aproximare a derivatei $f'''(x_0)$ cu restul de ordinul lui h^2 dezvoltînd $f(x_0 \pm h)$ şi $f(x_0 \pm 2h)$ după formula lui Taylor de ordin patru în jurul punctului x_0.

Utilizînd procedeul de extrapolare al lui Richarson să se obţină o formulă de aproximare a derivatei $f'''(x_0)$ cu restul de

ordinul lui h^4.

7. Funcţia $y = f(x)$ are un minim în intervalul $0,2 < x < 1,4$. Să se găsească abscisa punctului de minim, dacă se dă următorul tabel de valori ale lui f:

x	0,2	0,4	0,6	0,8	1,0	1,2	1,4
f(x)	2,10022	1,98730	1,90940	1,86672	1,85937	1,88737	1,95063

8. Să se calculeze $y'(1)$ şi $y''(1)$, unde $y = y(x)$ este dată de tabelul de valori

x	0,96	0,98	1,00	1,02	1,04
f(x)	0,7825361	0,7739332	0,7651977	0,7563321	0,7473390

(funcţia tabelată este funcţia Bessel de indice zero, $\mathcal{J}_0(x)$).

9. Să se găsească rădăcina primei derivate a funcţiei $y = \mathcal{J}_1(x)$ dată de tabelul de valori

x	1,80	1,82	1,84	1,86	1,88	1,90
$y_1(x)$	0,5815170	0,5817731	0,5818649	0,5817926	0,5815566	0,5811571

10. Funcţia $y = \mathcal{J}_n(x)$ este dată de tabelul de valori

x	85,00	85,01	85,02	85,03	85,04
$\mathcal{J}_n(x)$	0,0353878892	0,0346198696	0,0338490002	0,0330753467	0,0322989750

Să se afle numărul întreg şi pozitiv n, ştiind că funcţia este o soluţie a ecuaţiei lui Bessel de ordin n:

$$x^2y'' + xy' + (x^2 - n^2)y = 0 .$$

11. Funcţia $y = f(x)$ satisface ecuaţia diferenţială

$$xy'' + ay' + (x-b)y = 0 ,$$

unde a şi b sînt numere întregi. Să se determine constantele a şi b dîndu-se tabelul de valori

x	0,8	1,0	1,2	1,4	1,6	1,8	2,0	2,2
f(x)	1,73036	1,95532	2,19756	2,45693	2,73309	3,02549	3,33334	3,65563

Capitolul 8

METODE DE REZOLVARE NUMERICA A ECUATIILOR DIFERENTIALE ORDINARE SI CU DERIVATE PARTIALE

In acest capitol vor prezenta mai multe metode de rezolvare numerică a ecuaţiilor diferenţiale ordinare cît şi a unor clase de ecuaţii cu derivate parţiale.

Este bine cunoscut faptul că nu există metode generale de determinare explicită a soluţiei ecuaţiei diferenţiale $y'(x) = f(x,y(x))$, $x_0 \leq x \leq x_0 + \delta$, $y(x_0) = y_0$.

Cel mai simplu exemplu în acest sens este acela în care $f(x,y) = f(x)$ $(\forall) x \in [x_0, x_0+\delta]$, $(\forall)\ y \in R$ şi f este o funcţie continuă ale cărei primitive nu sînt funcţii elementare. Acelaşi fenomen are loc şi în cazul ecuaţiilor cu derivate parţiale de ordinul al doilea - chiar liniare - ori de cîte ori domeniul de definiţie al funcţiei necunoscute are o structură geometrică complicată.

Din aceste motive, în foarte multe situaţii practice, sîntem nevoiţi să lucrăm cu „soluţii aproximative", să punem în evidenţă metode cît mai eficiente de determinare a acestor soluţii cît şi metode de evaluare a „abaterii" acestor soluţii de la soluţia exactă.

Scopul acestui capitol este de a pune la îndemîna cititorului cîteva metode de acest tip, metode pe care practica le-a impus ca fiind cele mai simple şi uşor de implementat pe calculator.

§ 1. Noţiuni introductive

Fie $n \in N^*$ şi $\|\cdot\|: R^n \to R_+$ o normă pe R^n (vezi cap.I, § 4). Fie Ω o submulţime deschisă din R^n, $x_0 \in R$, $\delta > 0$ şi fie $f: [x_0, x_0+\delta] \times \Omega \longrightarrow R^n$ o funcţie continuă. Să considerăm problema lui Cauchy

$$(1.1)\qquad \begin{aligned} &\frac{dy}{dx}(x) = f(x, y(x)),\ x_0 \le x \le x_0 + \delta \\ &y(x_0) = y_0\ , \end{aligned}$$

unde $y_0 \in \Omega$ este data iniţială în $x = x_0$. Reamintim că (1.1) reprezintă scrierea vectorială a sistemului de ecuaţii diferenţiale cu condiţii iniţiale.

$$(1.2)\qquad \begin{cases} \dfrac{dy_1}{dx}(x) = f_1(x, y_1(x), y_2(x), \ldots, y_n(x)),\ x_0 \le x \le x_0+\delta \\ \dfrac{dy_2}{dx}(x) = f_2(x, y_1(x), y_2(x), \ldots, y_n(x)),\ x_0 \le x \le x_0+\delta \\ \vdots \\ \dfrac{dy_n}{dx}(x) = f_n(x, y_1(x), y_2(x), \ldots, y_n(x)),\ x_0 \le x \le x_0+\delta \\ y_1(x_0) = y_{01}, y_2(x_0) = y_{02}, \ldots, y_n(x_0) = y_{0n}\ , \end{cases}$$

legătura dintre (1.1) şi (1.2) fiind dată de

$$y(x) = \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix},\quad \frac{dy}{dx}(x) = \begin{pmatrix} \dfrac{dy_1}{dx}(x) \\ \dfrac{dy_2}{dx}(x) \\ \vdots \\ \dfrac{dy_n}{dx}(x) \end{pmatrix},\quad f(x,y) = \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \\ \vdots \\ f_n(x,y) \end{pmatrix}$$

pentru orice $(x,y) \in [x_0, x_0+\delta] \times \Omega$, iar $y_0 = \begin{pmatrix} y_{01} \\ y_{02} \\ \vdots \\ y_{0n} \end{pmatrix}$.

<u>Definiţia 1.1.</u> Spunem că funcţia $f: [x_0, x_0+\delta] \times \Omega \to R^n$ <u>satisface condiţia lui Lipschitz</u> pe mulţimea

$K \subset [x_0, x_0+\delta] \times \Omega$ (în raport cu variabila y) dacă există $L(K) > 0$ astfel încît

(1.3) $\|f(x,y) - f(x,z)\| \leq L(K)\|y-z\|$ $(\forall)(x,y),(x,z) \in K$.

Constanta $L(K)$ din (1.3) poartă numele de constantă Lipschitz corespunzătoare mulţimii K.

Observaţia 1.1. Dacă $f: [x_0, x_0+\delta] \times \Omega \longrightarrow R^n$ este o funcţie continuă cu proprietatea că pentru orice $y_0 \in \Omega$ există o vecinătate V a lui y_0 inclusă în Ω astfel încît f să satisfacă condiţia lui Lipschitz pe mulţimea $k = [x_0, x_0+\delta] \times V$, atunci (1.1) are proprietatea de existenţă şi unicitate locală. Aceasta înseamnă că pentru orice $y_0 \in \Omega$, există $\mu = \mu(y_0) \in (0, \delta]$ astfel încît (1.1) să admită o soluţie unică $y: [x_0, x_0+\mu] \longrightarrow \Omega$.

Dacă, în plus, $\Omega = R^n$ şi f satisface condiţia lui Lipschitz pe mulţimea $K = [x_0, x_0+\delta] \times R^n$, atunci pentru orice $y_0 \in \Omega$ problema (1.1) admite o soluţie unică definită pe întreg intervalul $[x_0, x_0+\delta]$. (Vezi Sabac [19, § 11, p.340].)

O condiţie suficientă, deosebit de utilă în aplicaţii, ca f să satisfacă condiţia lui Lipschitz pe mulţimile compacte din $[x_0, x_0+\delta] \times \Omega$ este dată de următoarea teoremă

Teorema 1.1 (a). Fie $h: [x_0, x_0+\delta] \longrightarrow R^n$ o funcţie de clasă C^1. Atunci

(1.4) $$\|h(x) - h(\bar{x})\|_1 \leq \sum_{i=1}^{n} \|h_i'\| \, |x-\bar{x}|$$

şi

(1.5) $$\|h(x) - h(\bar{x})\|_\infty \leq \max\{\|h_i'\| \; ; \; i = \overline{1,n}\} \, |x-\bar{x}|,$$

$(\forall)\ x, \bar{x} \in [x_0, x_0+\delta]$, unde $\|\cdot\|_1$ şi $\|\cdot\|_\infty$ sînt definite ca în cap. I, § 4, iar $\|h_i'\| = \sup\{|h'(s)| \; ; \; s \in [x_0, x_0+\delta]\}$.

(b). Fie $f: [x_0, x_0+\delta] \times \Omega \rightarrow R^n$ o funcţie de clasă C^1 şi fie $K \subset [x_0, x_0+\delta] \times \Omega$ o mulţime compactă. Atunci

(1.6) $\quad \|f(x,y)-f(x,z)\|_1 \le \max\left\{\sum_{j=1}^{n}\left\|\frac{\partial f_j}{\partial y_k}\right\|_K;\ k=\overline{1,n}\right\}\|y-z\|_1$

şi

(1.7) $\quad \|f(x,y)-f(x,z)\|_\infty \le \sum_{i=1}^{n}\max\left\{\left\|\frac{\partial f_j}{\partial y_i}\right\|_K;\ j=\overline{1,n}\right\}\|y-z\|_\infty$

pentru orice $(x,y),\ (x,z)\in K$, unde $\|\cdot\|_1$ şi $\|\cdot\|_\infty$ sînt definite ca în cap.I, § 4, iar

$$\left\|\frac{\partial f_j}{\partial y_i}\right\|_K = \sup\left\{\left|\frac{\partial f_j}{\partial y_i}(\xi,\eta)\right|;\ (\xi,\eta)\in K\right\}$$

pentru orice $i,j\in\{1,2,\dots,n\}$.

Dacă în plus, $\Omega = R^n$ şi $\frac{\partial f_j}{\partial y_i}$ sînt mărginite pe $[x_0,x_0+\delta]\times$ $\times R^m$, i,j $1,2,\dots,n$, atunci

(1.8) $\quad \|f(x,y)-f(x,z)\|_1 \le \max\left\{\sum_{j=1}^{n}\left\|\frac{\partial f_j}{\partial y_i}\right\|;\ i=\overline{1,n}\right\}\|y-z\|_1$

şi

(1.9) $\quad \|f(x,y)-f(x,z)\|_\infty \le \sum_{i=1}^{n}\max\left\{\left\|\frac{\partial f_j}{\partial y_i}\right\|;\ j=\overline{1,n}\right\}\|y-z\|_\infty$

pentru orice $(x,y),\ (x,z)\in[x_0,x_0+\delta]\times R^n$, unde

$$\left\|\frac{\partial f_j}{\partial y_i}\right\| = \sup\left\{\left|\frac{\partial f_j}{\partial y_i}(\xi,\eta)\right|;\ (\xi,\eta)\in[x_0,x_0+\delta]\times R^n\right\}.$$

Demonstraţie. Din Teorema lui Lagrange avem

$\|h(x)-h(\bar{x})\|_1 = \sum_{i=1}^{n}|h_i(x)-h_i(\bar{x})| = \sum_{i=1}^{n}|h_i'(\xi_i)||x-\bar{x}|$

cu $\xi_i\in(x,\bar{x})$ sau $\xi_i\in(\bar{x},x)$, $i=\overline{1,n}$. Cum $|h_i'(\xi_i)|\le\|h_i'\|$ pentru orice $i=\overline{1,n}$, rezultă

$$\|h(x)-h(\bar{x})\|_1 \le \sum_{i=1}^{n}\|h_i'\|\,|x-\bar{x}|$$

pentru orice $x,\bar{x}\in[x_0,x_0+\delta]$. Deci are loc (1.4).

Analog

$$\|h(x) - h(\bar{x})\|_\infty = \max\{|h_i(x) - h_i(\bar{x})| \; ; \; i = \overline{1,n}\} =$$

$$= \max\{|h_i'(\xi_i)||x-\bar{x}|; \quad i=\overline{1,n}\} = \max\{|h_i'(\xi_i)|; \; i=\overline{1,n}\}|x-\bar{x}|$$

unde $\xi_i \in (x,\bar{x})$, sau $\xi_i \in (\bar{x},x)$ pentru $i = \overline{1,n}$. Intrucit $|h_i'(\xi_i)| \leq \|h_i'\|$ pentru orice $i = \overline{1,n}$, rezultă

$$\|h(x) - h(\bar{x})\|_\infty \leq \max\{\|h_i'\|; \quad i=\overline{1,n}\}|x-\bar{x}|,$$

pentru orice $x, \bar{x} \in [x_o, x_o+\delta]$, ceea ce demonstrează (1.5).

Pentru a demonstra (1) şi (1.7) fie (x,y), $(x,z) \in K$ şi fie $W^o = (W_1, W_2, \ldots, W_n) = y-z$, $W^1 = (o, W_2, \ldots, W_n)$, $W^2 = (o,o,W_3,\ldots$ $\ldots, W_n), \ldots, W^{n-1} = (o,o,\ldots,o,W_n)$ şi $W^n = (o,o,\ldots,o)$.

Evident

$$(1.10) \quad \|f(x,y)-f(x,z)\|_1 = \|f(x,z+W^o)-f(x,z)\|_1 \leq \sum_{i=1}^{n} \|f(x,z+W^{i-1}) - f(x,z+W^i)\|_1 .$$

Din (1,4) deducem

$$\|f(x,z+W^{i-1}) - f(x,z+W^i)\|_1 \leq \sum_{j=1}^{n} \left\|\frac{\partial f_j}{\partial y_i}\right\|_K \cdot |W_i|$$

Cum $W_i = y_i - z_i$, $i=1,n$, din ultima inegalitate rezultă

$$\|f(x,z+W^{i-1})-f(x,z+W^i)\|_1 \leq \sum_{j=1}^{n} \left\|\frac{\partial f_j}{\partial y_i}\right\|_K |y_i - z_i| \leq$$

$$\leq \max\left\{\sum_{j=1}^{n} \left\|\frac{\partial f_j}{\partial y_k}\right\|_K ; \quad k = \overline{1,n}\right\} |y_i - z_i|$$

pentru orice $i = \overline{1,n}$. Din (1.10) şi ultima inegalitate obţinem (1.6).

Cu aceleaşi notaţii ca mai sus, avem

$$(1.11)\qquad \|f(x,y)-f(x,z)\|_\infty = \|f(x,z+W^0)-f(x,z)\|_\infty \leq$$

$$\leq \sum_{i=1}^{n} \|f(x,z+W^{i-1}) - f(x,z+W^i)\|_\infty .$$

Din (1.5) deducem

$$\|f(x,z+W^{i-1})-f(x,z+W^i)\|_\infty \leq \max\left\{\left\|\frac{\partial f_j}{\partial y_i}\right\|_K ;\quad j=\overline{1,n}\right\}\cdot |W_i| ,$$

Deoarece $|W_i| = |y_i - z_i| \leq \|y-z\|_\infty$, urmează

$$\|f(x,z+W^{i-1}) - f(x,z+W^i)\|_\infty \leq \max\left\{\left\|\frac{\partial f_j}{\partial y_i}\right\|_K ;\quad j=\overline{1,n}\right\} \|y-z\|_\infty .$$

Din ultima inegalitate şi din (1.11) rezultă (1.7).

Cum (1.8) şi (1.9) se demonstrează analog, nu mai dăm detalii.

Incheiem acest paragraf cu o inegalitate integrală deosebit de utilă.

Lema 1.1. (Inegalitatea lui Gronwall). Fie α, φ, ψ funcţii continue de la $[a,b]$ în R cu $\psi(x) \geq 0$ pentru orice $x \in [a,b]$. Dacă

$$(1.12)\qquad \alpha(x) \leq \varphi(x) + \int_a^x \psi(s)\alpha(s)ds, \text{ pentru orice } x \in [a,b],$$

atunci

$$(1.13)\qquad \alpha(x) \leq \varphi(x) + \int_a^x \varphi(s)e^{\int_s^x \psi(\tau)d\tau} ds, \text{ pentru orice } x \in [a,b].$$

Demonstraţie. Să notăm cu $\beta:[a,b] \longrightarrow R$ funcţia definită prin

$$(1.14)\qquad \beta(x) = \int_a^x \psi(s)\alpha(s)ds, \text{ pentru orice } x \in [a,b]$$

Evident $\beta'(x) = \psi(x)\alpha(x)$ şi cum $\psi(x) \geq 0$, din (1.12) rezultă

$$\beta'(x) \leq \psi(x)\varphi(x) + \psi(x)\beta(x), \text{ pentru orice } x \in [a,b].$$

Inmulţind în ambii membri această inegalitate cu $e^{-\int_a^x \psi(s)ds}$, obţinem

$$\frac{d}{dx}\left(\beta(x)e^{-\int_a^x \psi(s)ds}\right) \le \psi(x)\varphi(x)e^{-\int_a^x \psi(s)ds}$$, pentru orice $x\in[a,b]$. Integrînd inegalitatea de mai sus în ambii membri pe $[a,b]$, deducem

$$\beta(x) \le \int_a^x \psi(s)\varphi(s)e^{\int_s^x \psi(\tau)d\tau}ds \text{ , pentru orice } x\in[a,b].$$

Din (1.12), (1.14) și ultima inegalitate rezultă (1.13), ceea ce încheie demonstrația.

Corolarul 1.1. (Inegalitatea lui Bellman). Fie α și ψ două funcții continue pe $[a,b]$ cu $\psi(x)\geqslant 0$ pentru orice $x\in[a,b]$ și fie $M\in R$. Dacă

$$\alpha(x)\le M+\int_a^x \psi(s)\alpha(s)ds \text{ , pentru orice } x\in[a,b]$$

atunci

$$\alpha(x)\le M\,e^{\int_a^x \psi(s)ds} \text{ , pentru orice } x\in[a,b].$$

Demonstrație. Punînd $\varphi(x) = M$ pentru orice $x\in[a,b]$ sîntem în ipotezele Lemei 1.1. Deci

$$\alpha(x) = M + M\int_a^x \psi(s)e^{\int_s^x \psi(\tau)d\tau}ds =$$

$$= M - M\int_a^x \frac{d}{ds}\left(e^{\int_s^x \psi(\tau)d\tau}\right)ds = M - M\,e^{\int_s^x \psi(\tau)d\tau}\Big|_a^x =$$

$$= M - M + M\,e^{\int_a^x \psi(\tau)d\tau} = M\,e^{\int_a^x \psi(\tau)d\tau},$$

pentru orice $x\in[a,b]$.

§ 2. Soluții ε - aproximative

Incepem prin definirea conceptului de soluție ε - aproximativă pentru ecuația (1.1).

Definiția 2.1. Fie $\varepsilon>0$. O funcție $v_\varepsilon : [x_0,x_0+\mu]\to\Omega$ se numește soluție ε - aproximativă a problemei (1.1) pe $[x_0,x_0+\mu]\subset[x_0,x_0+\delta]$ în raport cu norma $\|\cdot\|$ pe R^n dacă

(i) y_ε este continuă pe $[x_0,x_0+\mu]$ şi $y_\varepsilon(x_0) = y_0$;

(ii) există o submulţime finită sau vidă F a intervalului $[x_0,x_0+\mu]$ astfel încît y_ε este de clasă C^1 pe $[x_0,x_0+\mu]-F$, iar y'_ε este mărginită pe $[x_0,x_0+\mu]-F$.

(iii) $\sup\{\| y'_\varepsilon(x) - f(x,y_\varepsilon(x))\| \ ; \ x\in[x_0,x_0+\mu]-F\} \leq \varepsilon$.

Cu alte cuvinte, o funcţie $y_\varepsilon:[x_0,x_0+\mu]\longrightarrow\Omega$ este o soluţie ε-aproximativă a problemei (1.1) pe $[x_0,x_0+\mu]$ dacă y_ε satisface (i) şi (ii) din Definiţia 2.1 şi în plus

(2.1) $$y'_\varepsilon(x) = f(x,y_\varepsilon(x)) + g_\varepsilon(x), \quad \text{pentru orice} \quad x\in[x_0+\zeta+\mu]-F$$
$$y_\varepsilon(x_0) = y_0 ,$$

unde $g_\varepsilon:[x_0,x_0+\mu]\longrightarrow R^n$ este o funcţie continuă şi mărginită pe $[x_0,x_0+\mu]-F$ care satisface

(2.2) $$\sup\{\|g_\varepsilon(x)\| \ ; \ x\in[x_0,x_0+\mu]-F\}\leq\varepsilon.$$

Evident, soluţia y a problemei (1.1) este soluţie ε-aproximativă pe intervalul ei de existenţă pentru orice $\varepsilon > 0$.

In cele ce urmează ne vom ocupa de două probleme fundamentale referitoare la soluţiile ε-aproximative ale ecuaţiei (1.1).

Prima problemă este aceea a determinării unor condiţii suficiente asupra funcţiei f astfel încît pentru orice $\varepsilon > 0$ problema (1.1) să admită cel puţin o soluţie ε-aproximativă care să poată fi construită efectiv.

Cea de a doua problemă constă în determinarea „abaterii" unei soluţii ε-aproximative de la soluţia „exactă".

In acest paragraf ne vom ocupa în exclusivitate de cea de a doua problemă.

Definiţia 2.2. Fie $y:[x_0,x_0+\mu]\rightarrow\Omega$ soluţia problemei (1.1) şi fie $y_\varepsilon:[x_0,x_0+\mu]\rightarrow\Omega$ o soluţie ε-aproximativă a aceleiaşi probleme. Prin eroarea soluţiei ε-aproximative y_ε înţelegem numărul pozitiv $e(y_\varepsilon)$ definit prin

$$e(y_\varepsilon) = \sup\{\|y(x) - y_\varepsilon(x)\| \; ; \; x \in [x_0, x_0+\mu]\}.$$

Prin majorant al erorii soluţiei ε-aproximative y_ε înţelegem orice număr pozitiv $m(y_\varepsilon)$ care satisface

$$e(y_\varepsilon) \leq m(y_\varepsilon).$$

Datorită lipsei de informaţii asupra soluţiei „exacte" y, practic este imposibil să se determine eroarea $e(y_\varepsilon)$. Din acest motiv, problema centrală în acest context constă în <u>determinarea unui majorant al erorii $m(y_\varepsilon)$ cît mai apropiat de $e(y_\varepsilon)$</u>.

Rezultatul central în această privinţă este:

<u>Teorema 2.1.</u> <u>Fie $f: [x_0, x_0+\delta] \times \Omega \to R^n$ o funcţie de clasă C^1, fie $y: [x_0, x_0+\mu] \to \Omega$ soluţia problemei (1.1) cu $\mu \in (0, \delta]$, fie $y_\varepsilon: [x_0, x_0+\mu] \to \Omega$ o soluţie ε-aproximativă a problemei (1.1) în raport cu norma $\|\cdot\|_1$ pe R^n şi fie $z_\varepsilon: [x_0, x_0+\mu] \to \Omega$ o soluţie ε-aproximativă a problemei (1.1) în raport cu norma $\|\cdot\|_\infty$ pe R^n. Dacă se cunoaşte o submulţime compactă D a lui Ω astfel încît $y(x) \in D$, $y_\varepsilon(x) \in D$ şi $z_\varepsilon(x) \in D$ pentru orice $x \in [x_0, x_0+\mu]$ atunci un majorant al erorii pentru y_ε (z_ε) este</u>

$$m(y_\varepsilon) = \mu e^{L\mu} \cdot \varepsilon \tag{2.3}$$

$$m(z\;) = \mu e^{\tilde{M}\mu} \cdot \varepsilon, \tag{2.4}$$

<u>unde</u>

$$L = \max\left\{ \sum_{j=1}^{n} \left\| \frac{\partial f_j}{\partial y_k} \right\|_D \; ; \quad k = \overline{1,n} \right\}$$

$$\tilde{M} = \sum_{i=1}^{n} \max\left\{ \left\| \frac{\partial f_i}{\partial y_j} \right\|_D \; ; \quad j = \overline{1,n} \right\}$$

<u>iar</u>

$$\left\| \frac{\partial f_j}{\partial y_k} \right\|_D = \sup\left\{ \left| \frac{\partial f_j}{\partial y_k}(\xi, \eta) \right| ; \; (\xi, \eta) \in [x_0, x_0+\mu] \times D \right\}.$$

<u>Dacă, în plus, $\frac{\partial f_i}{\partial y_j}$ sînt mărginite pe $[x_0, x_0+\mu] \times \Omega$</u>

atunci cunoașterea mulțimii D de mai sus nu mai este necesară, (2.3) și (2.4) avînd loc cu

$$L = \max\left\{\sum_{j=1}^{n} \left\|\frac{\partial f_j}{\partial y_k}\right\| ;\quad k = \overline{1,n}\right\}$$

$$\tilde{M} = \sum_{i=1}^{n} \max\left\{\left\|\frac{\partial f_j}{\partial y_i}\right\| ;\quad j = \overline{1,n}\right\}$$

unde

$$\left\|\frac{\partial f_j}{\partial y_k}\right\| = \sup\left\{\left|\frac{\partial f_j}{\partial y_k}(\xi,\eta)\right| ;\quad (\xi,\eta)\in [x_0,x_0+\mu]\times\Omega\right\}.$$

Observația 2.1. Dificultatea principală în aplicarea primei părți a Teoremei 2.1 constă în determinarea submulțimii compacte D astfel încît $y(x)\in D$ pentru orice $x\in[x_0,x_0+\mu]$ și aceasta deoarece y nu este cunoscută. Cu toate acestea, în numeroase situații concrete cînd y are o semnificație fizică bine precizată putem stabili apriori niște „limite" între care y poate lua valori, fapt care conduce la determinarea mulțimii D.

Demonstrația Teoremei 2.1. Din (1.1) și (2.1) rezultă

$$y'_\varepsilon(x) - y'(x) = f(x,y_\varepsilon(x)) - f(x,y(x)) + g_\varepsilon(x)$$

pentru orice $x\in[x_0,x_0+\mu]-F$, unde F este mulțimea finită sau vidă din Definiția 2.1. Pentru a fixa ideile să presupunem că $F=\{x_j;\ j=\overline{1,k}\}$, unde $x_0\le x_1\le \dots \le x_k\le x_0+\mu = x_{k+1}$. Fie $x\in[x_0,x_0+\mu]$ și fie $j\in\{0,1,\dots,k\}$ cu $x_j\le x\le x_{j+1}$. Avem atunci

$$\sum_{p=0}^{j-1}\int_{x_p}^{x_{p+1}}[y'_\varepsilon(s)-y'(s)]ds + \int_{x_j}^{x}[y'_\varepsilon(s)-y'(s)]ds =$$

$$= \int_{x_0}^{x}[f(s,y_\varepsilon(s)) - f(s,y(s))]ds + \int_{x_0}^{x} g_\varepsilon(s)ds.$$

Din această egalitate deducem

$$\sum_{p=0}^{j-1}[v_\varepsilon(s)-v(s)]\Big|_{x_p}^{x_{p+1}}+[v_\varepsilon(s)-v(s)]\Big|_{x_j}^{x}=$$

$$=\int_{x_0}^{x}[f(s,v_\varepsilon(s))-f(s,v(s))]ds+\int_{x_0}^{x}g_\varepsilon(s)ds,$$

de unde rezultă

$$v_\varepsilon(x)-v(x)=\int_{x_0}^{x}[f(s,v_\varepsilon(s))-f(s,v(s))]ds+\int_{x_0}^{x}g_\varepsilon(s)ds.$$

De aici obţinem

$$(2.5)\qquad \|v_\varepsilon(x)-v(x)\|_1\leq\int_{x_0}^{x}\|g_\varepsilon(s)\|_1ds+\int_{x_0}^{x}\|f(s,v_\varepsilon(s))-f(s,v(s))\|_1ds$$

pentru orice $x\in[x_0,x_0+\mu]$.

Ţinînd cont că $v(s)\in D$ şi $v_\varepsilon(s)\in D$ pentru orice $s\in[x_0,x_0+\mu]$ din (1.6) în Teorema 1.1 urmează

$$\|f(s,v_\varepsilon(s))-f(s,v(s))\|_1\leq L\|v_\varepsilon(s)-v(s)\|_1$$

pentru orice $s\in[x_0,x_0+\mu]$ unde L este definit ca în enunţ. Această inegalitate combinată cu (2.2) şi cu (2.5) conduce la

$$\|v_\varepsilon(x)-v(x)\|_1\leq(x-x_0)\varepsilon+\int_{x_0}^{x}L\|v_\varepsilon(s)-v(s)\|_1ds$$

pentru orice $x\in[x_0,x_0+\mu]$. Intrucît $x-x_0\leq\mu$ şi $\varepsilon>0$, avem

$$(2.6)\qquad \|v_\varepsilon(x)-v(x)\|_1\leq\mu\cdot\varepsilon+\int_{x_0}^{x}L\|v_\varepsilon(s)-v(s)\|_1ds$$

pentru orice $x\in[x_0,x_0+\mu]$.

Luînd $\alpha(x)=\|v_\varepsilon(x)-v(x)\|_1$, $\psi(x)=L$ şi $M=\mu\cdot\varepsilon$ în Corolarul 1.1, şi ţinînd cont de (2.6) obţinem

$$\|v_\varepsilon(x)-v(x)\|_1\leq\mu\cdot\varepsilon\cdot e^{\int_{x_0}^{x}L\,ds}=\mu\cdot\varepsilon\cdot e^{L(x-x_0)}\leq\mu\cdot\varepsilon\cdot e^{L\mu}$$

pentru orice $x\in[x_0,x_0+\mu]$, ceea ce demonstrează (2.3).

Cum verificarea relaţiei (2.4) cît şi a părţii a doua a Teoremei 2.1 se face analog, le lăsăm în seama cititorului.

Incheiem acest paragraf cu o consecință deosebit de utilă a Teoremei 2.1.

Teorema 2.2. Fie $f:[x_0,x_0+\delta]\times\Omega\to R^n$ o funcție de clasă C^1 și fie $y:[x_0,x_0+\mu]\to\Omega$ soluția problemei (1.1). Dacă pentru orice $k\in N^*$ există cel puțin o soluție $1/k$ - aproximativă $v_{1/k}:[x_0,x_0+\mu]\to\Omega$ a problemei (1.1) în raport cu norma $\|\cdot\|_1$ pe R^n și dacă există o submulțime compactă D a lui Ω astfel încît $y(x)\in D$ și $v_{1/k}(x)\in D$ pentru orice $k\in N^*$ și orice $x\in[x_0,x_0+\mu]$, atunci

(2.7) $$\lim_{k\to+\infty} v_{1/k}(x) = y(x) \text{ uniform pe } [x_0,x_0+\mu].$$

Dacă, în plus $\frac{\partial f_j}{\partial y_i}$ sînt mărginite pe $[x_0,x_0+\mu]\times\Omega$, atunci cunoașterea submulțimii D nu mai este necesară, (2.7) avînd de asemenea loc.

Observația 2.2. Inlocuind $\|\cdot\|_1$ cu $\|\cdot\|_\infty$ în Teorema 2.2 se obține un rezultat analog. Lăsăm în seama cititorului enunțul și demonstrația acestui rezultat.

Demonstrația Teoremei 2.2. Din Teorema 2.1 combinată cu Definiția 2.2 rezultă

$$\sup\{\|y(x)-v_{1/k}(x)\|_1;\ \ x\in[x_0,x_0+\mu]\}\le\mu\cdot e^{L\mu}\cdot 1/k$$

pentru orice $k\in N^*$, inegalitate ce demonstrează (2.7).

Partea a doua a Teoremei 2.2 se demonstrează analog.

§ 3. Metoda lui Euler (a liniilor poligonale) de rezolvare numerică a ecuațiilor diferențiale de ordinul întîi

Fie $f:[x_0,x_0+\delta]\times R^n\to R^n$ o funcție de clasă C^1 și să considerăm problema lui Cauchy

$$\frac{dy}{dx}(x) = f(x,y(x)), \quad x_0 \le x \le x_0 + \delta$$

(3.1)

$$y(x_0) = y_0$$

unde $y_0 \in R^n$ este data inițială în $x = x_0$.

Fie $y: [x_0, x_0+\mu] \longrightarrow R^n$ unica soluție locală a problemei (3.1) cu $\mu \in (0,\delta]$ și fie $\varepsilon > 0$. In cele ce urmează ne propunem să punem în evidență un procedeu de construcție a unei soluții ε-aproximative $y_\varepsilon : [x_0, x_0+\mu] \longrightarrow R^n$ pentru problema (3.1) în raport cu norma $\|\cdot\|_1$ pe R^n. Cazul cînd $\|\cdot\|_1$ se înlocuiește cu $\|\cdot\|_\infty$ se tratează analog. In acest scop fie $m \in N^*$ al cărui ordin de mărime - depinzînd de $\varepsilon > 0$ - va fi precizat ulterior, fie $h = \mu/m$ și să considerăm partiția (P_h) de pas h a intervalului $[x_0, x_0+\mu]$

$$(P_h): \; x_0 < x_1 < \dots < x_m = x_0+\mu \,, \quad x_{i+1}-x_i = h, \; i=\overline{0,m-1}.$$

Intr-o primă etapă vom determina valorile $y_0, y_1, \dots, y_m$ ale soluției ε-aproximative în nodurile $x_0, x_1, \dots, x_m$, după care vom defini soluția ε-aproximativă ca fiind acea funcție al cărei grafic este linia poligonală ce unește (x_i, y_i) cu (x_{i+1}, y_{i+1}) pentru $i = \overline{0,m-1}$.

Algoritmul propus pentru calculul valorilor $y_i = y_\varepsilon(x_i)$, $i = \overline{0,m-1}$ este următorul

(3.2) $$y_{i+1} = y_i + h\, f(x_i, y_i), \quad i = \overline{0,m-1}$$

unde y_0 este dat de (3.1).

Acest algoritm este sugerat de cazul unidimensional (n=1), cînd pentru soluția exactă y a problemei (3.1) avem (conform teoremei lui Lagrange):

$$y(x_{i+1}) = y(x_i) + y'(\xi_i)(x_{i+1}-x_i), \quad i = \overline{0,m-1},$$

unde $\xi_i \in (x_i, x_{i+1})$. Atunci, pentru h „suficient de mic", putem aproxima $y'(\xi_i)$ cu $y'(x_i) = f(x_i, y_i)$, ajungînd astfel la (3.2).

Cu valorile $y_0, y_1, \dots, y_m$ determinate prin (3.2) putem construi funcţia $y_h : [x_0, x_0+\mu] \to R^n$ prin

(3.3) $y_h(x) = y_i + (x - x_i) f(x_i, y_i)$, pentru $x \in [x_i, x_{i+1})$, $i = \overline{o,m-2}$ şi $x \in [x_{m-1}, x_m]$.

Se observă cu uşurinţă că y_h este continuă pe $[x_0, x_0+\mu]$, $y_h(x_i) = y_i$ pentru $i = \overline{o,m}$ şi că y_h este derivabilă pe (x_i, x_{i+1}), $i = \overline{o,m-1}$. In plus, avem

(3.4) $y_h'(x) = f(x_i, y_i)$ pentru $x \in (x_i, x_{i+1})$, $i = \overline{o,m-1}$.

In cele ce urmează vom arăta că dacă $\mu \in (o, \delta]$ este „destul de mic", pentru $m \in N^*$ „suficient de mare", funcţia y_h este o soluţie ε-aproximativă a problemei (3.1) pe $[x_0, x_0+\mu]$ în raport cu norma $\|\cdot\|_1$.

Pentru aceasta să definim funcţia $a_m : [x_0, x_0+\mu] \to [x_0, x_0+\mu]$,

(3.5) $$a_m(x) = \begin{cases} x_i & \text{dacă } x \in [x_i, x_{i+1}), \quad i = \overline{o,m-2} \\ x_{m-1} & \text{dacă } x \in [x_{m-1}, x_m]. \end{cases}$$

Să observăm că (3.4) poate fi rescrisă echivalent sub forma

$y_h'(x) = f(a_m(x), y_h(a_m(x)))$ pentru orice $x \in [x_0, x_0+\mu]$, $x \neq x_i$, $i = \overline{1,m-1}$.

Notînd cu $g_m : [x_0, x_0+\mu] \to R^n$ funcţia definită prin

$g_m(x) = f(a_m(x), y_h(a_m(x))) - f(x, y_h(x))$ pentru $x \in [x_0, x_0+\mu]$

egalitatea precedentă poate fi pusă sub forma

(3.6) $y_h'(x) = f(x, y_h(x)) + g_m(x)$, pentru $x \in [x_0, x_0+\mu]$, $x \neq x_i$, $i = \overline{1,m-1}$.

Pentru a demonstra că y_h este soluţie ε-aproximativă a problemei (3.1) în norma $\|\cdot\|_1$ de îndată ce m este „destul de mare", este suficient să arătăm că

(3.7) $$\lim_{m\to+\infty} \sup\left\{\|g_m(x)\|_1 \;;\; x\in[x_o,x_o+\mu]\right\} = 0.$$

vezi (2.1) şi (2.2).

In vederea stabilirii relaţiei (3.7) vom demonstra întîi următoarea lemă

<u>Lema 3.1. Dacă</u> $\mu\in(0,\delta]$, $r>0$ <u>şi</u> $M>0$ <u>satisfac</u>

(3.8) $\|f(x,y)\|_1\le M$ <u>pentru orice</u> $x\in[x_o,x_o+\mu]$ <u>şi orice</u> $y\in R^n$ <u>cu</u> $\|y-y_o\|_1\le r$ <u>şi</u>

(3.9) $$\mu\cdot M\le r ,$$

<u>atunci pentru orice</u> $m\in N^*$, <u>funcţia</u> y_h <u>definită de (3.3) satisface</u>

(3.10) $\|y_h(x)-y_o\|_1\le r$ <u>pentru orice</u> $x\in[x_o,x_o+\mu]$.

<u>Demonstraţie</u>. Din (3.8) avem

$$\|f(x_o,y_o)\|_1\le M$$

şi atunci

$$\|y_h(x)-y_o\|_1\le|x-x_o|\cdot\|f(x_o,y_o)\|_1\le\frac{\mu}{m}\cdot M<\mu\cdot M\le r$$

pentru orice $x\in[x_o,x_1]$. Pentru $x=x_1$ obţinem

$$\|y_1-y_o\|_1\le\frac{\mu}{m}\cdot M<r ,$$

şi din (3.8) urmează

$$\|f(x_1,y_1)\|_1\le M$$

In consecinţă

$$\|y_h(x)-y_o\|_1\le\|y_1+(x-x_1)f(x_1,y_1)-y_o\|_1\le\|y_1-y_o\|_1+$$

$$+\frac{\mu}{m}\|f(x_1,y_1)\|\le\frac{\mu}{m}\cdot M+\frac{\mu}{m}\cdot M<\mu M\le r,$$

pentru orice $x \in [x_1, x_2]$. In particular

$$\|v_2 - v_0\|_1 \leq \frac{2\mu}{m} M < r.$$

Repetînd raţionamentul de mai sus pe $[x_2, x_3]$, $[x_3, x_4]$ş.a.m.d., deducem că

$$\|v_h(x) - v_0\|_1 \leq \frac{i\mu}{m} \cdot M \leq r$$

pentru orice $i = \overline{1,m}$ şi $x \in [x_{i-1}, x_i]$, inegalitate ce implică (3.10).

Observaţia 3.1. Putem întotdeauna alege numerele $\mu \in (o, \delta]$, $r > o$ şi $M > o$ care să satisfacă împreună (3.8) şi (3.9).

Intr-adevăr, dacă $\tilde{\mu} \in (o, \delta]$ şi $r > o$ sînt două numere date, mulţimea $\{(x,v) \in [x_o, x_o + \tilde{\mu}] \times R^n \ ; \ x \in [x_o, x_o + \tilde{\mu}], \|v - v_o\|_1 \leq r\}$ este compactă în R^{n+1} şi cum f este continuă (fiind de clasă C^1), f este mărginită pe această mulţime. Există deci $M > o$ astfel încît

$\|f(x,v)\|_1 \leq M$ pentru orice $x \in [x_o, x_o + \tilde{\mu}]$ şi $v \in R^n$ cu $\|v - v_o\|_1 \leq r$. Luînd acum un număr $\mu \in (o, \tilde{\mu}]$ astfel încît

$$\mu M \leq r,$$

observăm că $\mu > o$, $r > o$ şi $M > o$, definite ca mai sus, satisfac împreună (3.8) şi (3.9).

Din cele arătate mai sus rezultă că toate consideraţiile care vor urma se vor referi numai la cazul în care $\mu \in (o, \delta]$ este „suficient de mic".

Teorema 3.1. Dacă $\mu \in (o, \delta]$, $r > o$ şi $M > o$ satisfac (3.8) şi (3.9), atunci are loc (3.7).

Demonstraţie. Avem

$$\|g_m(x)\|_1 = \|f(a_m(x), v_h(a_m(x))) - f(x, v_h(x))\|_1 \leq$$
$$\leq \|f(a_m(x), v_h(a_m(x))) - f(a_m(x), v_h(x))\|_1 +$$
$$+ \|f(a_m(x), v_h(x)) - f(x, v_h(x))\|_1 ,$$

pentru orice $x \in [x_o, x_o + \mu]$.

Fie $K = \{(x,y) \in [x_o,x_o+\delta] \times R^n;\ x \in [x_o,x_o+\mu],\ \|y-y_o\|_1 \leq r\}$.

Intrucît K este compactă şi din Lema 3.1 avem

$$(a_m(x),\ y_h(a_m(x))),\quad (x,y_h(x)),\ (a_m(x),y_h(x)) \in K\ ,$$

din Teorema 1.1 deducem

(3.11) $\|\varepsilon_m(x)\|_1 \leq L(K)\,\|y_h(a_m(x))- y_h(x)\|_1+ M(K)|a_m(x)- x|$

pentru orice $x \in [x_o,x_o+\mu]$, unde

(3.12) $$L(K) = \max\left\{ \sum_{j=1}^{n} \left\|\frac{\partial f_j}{\partial y_k}\right\|_K ;\quad k = \overline{1,n}\right\}$$

iar

(3.13) $$M(K) = \sum_{i=1}^{n} \left\|\frac{\partial f_i}{\partial x}\right\|_K\ ,$$

$$\left\|\frac{\partial f_j}{\partial y_k}\right\|_K = \sup\left\{\left|\frac{\partial f_j}{\partial y_k}(\xi,\eta)\right| ;\ (\xi,\eta) \in K\right\},$$

$$\left\|\frac{\partial f_i}{\partial x}\right\|_K = \sup\left\{\left|\frac{\partial f_i}{\partial x}(\xi,\eta)\right|;\ \ (\xi,\eta) \in K\right\},$$

$i,j,k \in \{1,2,\dots,n\}$.

Observînd că

$$\|y_h(a_m(x))- y_h(x)\|_1 = \|y_i- y_i-(x-x_i)f(x_i,y_i)\|_1$$
$$\leq\ M(x-x_i) \leq M\ h = M\frac{\mu}{m}$$

şi că

$$|a_m(x) - x| \leq |x-x_i| \leq h = \frac{\mu}{m}$$

pentru orice $x \in [x_i,x_{i+1}]$, $i = \overline{o,m-1}$, din (3.11) obţinem

(3.15) $$\|\varepsilon_m(x)\|_1 \leq [L(K).M + M(K)]\frac{\mu}{m}$$

pentru orice $x \in [x_o,x_o+\mu]$ şi orice $m \in N^*$.

Evident, ultima inegalitate implică (3.7), ceea ce încheie demonstraţia Teoremei 3.1.

<u>Observaţia 3.2</u>. Din Teorema 3.1 rezultă că pentru $m \in N^*$

pentru care $[L(K).M + M(K)]\cdot\frac{\mu}{m} \leq \varepsilon$, funcţia y_h este o soluţie ε -aproximativă a ecuaţiei (3.1). Din Teorema 2.1 deducem

$$(3.16) \qquad \|y_h(x) - y(x)\|_1 \leq [L(K)M + M(K)]\cdot\frac{\mu^2}{m}\, e^{L(K)\mu}$$

pentru orice $x \in [x_0, x_0+\mu]$, unde L(K) şi M(K) sînt daţi de (3.12) şi (3.13) respectiv, iar $M = \sup\{\|f(x,y)\|_1;\ (x,y)\in K\}$.

Formula (3.16) ne furnizează o metodă concretă de evaluare a erorii în cazul metodei lui Euler.

In cazul în care sistemul (3.1) este liniar şi omogen, adică $f(x,y) = Ay$ pentru orice $(x,y) \in [x_0, x_0+\delta] \times R^n$, unde

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \in \mathcal{M}_m(R),$$

iar $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, formula (3.16) capătă o formă mai simplă. Mai precis, avem

$$(3.17) \qquad \|y_h(x) - y(x)\|_1 \leq [r + \|y_0\|_1]\, \|A\|_c^2\, \frac{\mu}{m}\, e^{\|A\|_c \mu}$$

pentru orice $x \in [x_0, x_0+\mu]$, unde $r > o$ este ca în Teorema 3.1, iar $\|A\|_c$ este norma matricii A pe coloane. Vezi cap.I, (5.12).

Subliniem că rezolvarea unui sistem liniar de n ecuaţii diferenţiale cu n funcţii necunoscute, presupune rezolvarea unei ecuaţia polinomiale de grad n. Ori, pentru $n \geq 5$, nu există formule de rezolvare a ecuaţiilor polinomiale. De aceea, chiar în cazul liniar, problema determinării soluţiilor ε -aproximative şi a evaluării erorii este de un mare interes practic.

Incheiem acest paragraf cu un exemplu care scoate în evidenţă faptul că, în cazul neliniar, metoda lui Euler nu este eficientă din punct de vedere al volumului de calcul, deoarece

viteza de convergență a șirului de soluții ε-aproximative la soluția exactă este foarte mică.

Exemplu 3.1. Să considerăm ecuația

$$(3.18)\qquad \begin{aligned} y'(x) &= x^2 + y^2(x) \quad , \quad 0 \leq x \leq \mu \\ y(0) &= 1 \end{aligned}$$

Ne propunem să determinăm $\mu > 0$ și $m \in N^*$ astfel încît y_h definită de (3.3) să fie o soluție 10^{-2} aproximativă a ecuației (3.13) pe $[0, \mu]$.

Să observăm că în cazul de față $f(x,y) = x^2 + y^2$ pentru orice $x \in [0,\mu]$ și $y \in R$. Luînd $\mu > 0$ și $r = 1$ observăm că

$$|f(x,y)| = x^2 + y^2 \leq \mu^2 + y^2$$

pentru orice $x \in [0,\mu]$ și $y \in R$. Dacă $|y-1| \leq 1$, atunci $|y| \leq |y-1| + 1 \leq 2$ și ca atare

$$|f(x,y)| \leq \mu^2 + 4 = M$$

Punînd condiția ca $\mu . M \leq 1$ obținem

$$\mu^3 + 4\mu - 1 \leq 0,$$

inegalitate care este satisfăcută pentru $\mu = 0,2$.

Avem atunci

$$M = 4,04$$

$$L(K) = \left\| \frac{\partial f}{\partial y} \right\| = 4$$

$$M(K) = \left\| \frac{\partial f}{\partial x} \right\| = 0,4 \ .$$

In consecință, din (3.16), urmează

$$|y_h(x) - y(x)| \leq (4.4,04+0,4). \frac{0,04}{m} e^{4.0,2} \leq$$

$$\leq \frac{0,6624}{m} e^{0,8} \leq \frac{0,6624}{m} .3 = \frac{1,9872}{m} < \frac{1,99}{m} .$$

Rezultă atunci că y_h aproximează y pe $[0, 0,2]$ cu o eroare mai mică decît 10^{-2} dacă $\frac{1,99}{m} \leq 10^{-2}$, ceea ce este echivalent cu $m \geq 199$.

Acest exemplu evidențiază o deficiență majoră a metodei lui Euler și anume viteza de convergență foarte mică a șirului de soluții ε-aproximative y_h la y.

§ 4. Metoda Runge - Kutta de rezolvare numerică a ecuațiilor diferențiale de ordinul întîi

Fie $f:[x_o,x_o+\delta]\times R\to R$ o funcție de clasă C^m și să considerăm problema lui Cauchy

$$(4.1)\quad \begin{cases} \frac{dy}{dx}(x) = f(x,y(x)), & x_o \leq x \leq x+\delta \\ y(x_o) = y_o & \end{cases}$$

unde $y_o \in R$ este data inițială în $x = x_o$.

Fie $y:[x_o,x_o+\mu]\to R$ unica soluție locală a problemei (4.1) și fie (P): $x_o < x_1 < \ldots < x_n = x_o+\mu$ o diviziune a intervalului $[x_o,x_o+\mu]$. Dacă f are o expresie analitică „simplă", atunci, din (4.1) putem calcula $y''(x_o)$, $y'''(x_o),\ldots,y^{(m+1)}(x_o)$. Mai precis

$$y''(x_o) = \left(\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f\right)_o$$

$$y'''(x_o)= \left[\frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x\partial y} f+ \frac{\partial f}{\partial y}\frac{\partial f}{\partial x} + \frac{\partial^2 f}{\partial y^2} f^2+\left(\frac{\partial f}{\partial y}\right)^2 f\right]_o,$$

etc., unde $(\quad)_o$ sau $[\quad]_o$ semnifică faptul că funcția dintre paranteze este calculată în (x_o,y_o).

Din formula lui Taylor, avem atunci

$$y(x_1)=y(x_o)+\frac{y'(x_o)}{1!}(x_1-x_o)+\ldots+\frac{y^{(m)}(x_o)}{m!}(x_1-x_o)^m + \frac{y^{(m+1)}(\xi)}{(m+1)!}(x_1-x_o)^{m+1},$$

unde $\xi \in (x_o,x_1)$. Renunțînd la ultimul termen al dezvoltării, putem găsi valoarea aproximativă a lui $y(x_1)$.

Pornind de la $y(x_1)$ și repetînd procedeul de mai sus putem determina $y(x_2)$, ș.a.m.d.

In acest mod, pentru orice $\varepsilon > 0$, putem găsi o diviziune (P) a lui $[x_0, x_0+\mu]$ cu norma suficient de mică, astfel încît funcţia al cărei grafic este linia poligonală ce uneşte $(x_i, y(x_i))$ cu $(x_{i+1}, y(x_{i+1}))$, $i=\overline{0,n-1}$ să fie o soluţie ε-aproximativă a problemei (4.1).

Din păcate, această metodă nu este eficace decît într-un număr relativ mic de situaţii, întrucît presupune un volum foarte mare de calcule - adeseori extrem de greoaie - pentru determinarea valorilor $y''(x_0), y'''(x_0), \dots, y^{(m)}(x_0)$.

Din acest motiv s-a încercat să se pună în evidenţă o metodă de calcul aproximativ al valorii $y_{i+1} = y(x_i)$ pornind de la $y_i = y(x_i)$, care să evite determinarea valorilor $y''(x_i)$, $y'''(x_i), \dots, y^{(m)}(x_i)$ şi care să fie cît mai exactă cu putinţă.

Descriem în continuare o astfel de metodă. Să presupunem că $x_{i+1} - x_i = h$ constant - pentru orice $i = \overline{0,n-1}$. Cu notaţiile $y_i = y(x_i)$, $y'_i = y'(x_i)$, $y''_i = \dots, \dots$, avem din formula lui Taylor

$$y_{i+1} = y_i + \frac{h}{1!} y'_i + \frac{h^2}{2!} y''_i + \dots + \frac{h^m}{m!} y_i^{(m)} + \frac{h^{m+1}}{(m+1)!} y^{(m+1)}(\xi),$$

unde $\xi \in (x_i, x_{i+1})$.

Ideea de bază a metodei Runge - Kutta constă în înlocuirea expresiei din membrul drept al egalităţii de mai sus printr-o funcţie A_i - de variabilă h - ale cărei valori să poată fi determinate cît mai uşor şi astfel încît eroarea comisă prin această înlocuire să fie cît mai mică. Se caută deci y_{i+1} de forma

$$y_{i+1} = y_i + A_i(h), \qquad i = \overline{0,n-1}. \tag{4.2}$$

Intrucît funcţia A_i se determină la fel pentru orice $i=\overline{0,n-1}$, în continuare ne vom limita mai la cazul $i=0$. Formula (4.2) se scrie:

$$y_1 = y_0 + A_0(h)$$

Căutăm A_o de forma

(4.3) $$A_o(h) = P_{r1}\,\sigma_1(h) + P_{r2}\,\sigma_2(h) + \dots + P_{rr}\,\sigma_r(h)$$

unde $P_{r1}, P_{r2}, \dots, P_{rr}$ sînt constante, iar σ_i sînt definite prin

(4.4) $$\sigma_i(h) = h \cdot f(\xi_i, \eta_i), \quad i = \overline{1,r}\ ,$$

unde

(4.5) $$\begin{cases} \xi_i = x_o + \alpha_i h\ , \quad \alpha_1 = o\ , \quad i = \overline{2,r} \\ \eta_i = y_o + \beta_{i1}\sigma_1(h) + \beta_{i2}\sigma_2(h) + \dots + \beta_{i,i-1}\sigma_{i-1}(h), \quad i = \overline{\cdot,r} \end{cases}$$

α_i şi β_{ij} fiind, de asemenea, constante.

Eroarea comisă prin utilizarea lui $A_o(h)$ în locul lu $\Delta y_o = y_1 - y_o$ este dată de funcţia

$$\varphi_r(h) = \Delta y_o - A_o(h) = y(x_o + h) - y(x_o) - A_o(h).$$

Evident $\varphi_r(o) = o$. Este uşor de văzut că, dacă $\varphi_r'(o) = \varphi_r''(o) = \dots = \varphi_r^{(s)}(o) = o$, atunci Δy_o şi $A_o(h)$ au coeficieţii puterilor lui h din dezvoltările Taylor egali pînă la ordinul s.

In consecinţă, pentru determinarea constantelor P_{ri}, α_i şi β_{ij} se va impune condiţia ca un număr cît mai mare de derivate ale funcţiei φ_r să fie nule în h = o.

Intr-adevăr, dacă $\varphi_r'(o) = \varphi_r''(o) = \dots = \varphi_r^{(s)}(o)$ şi $\varphi_h^{(s+1)}(o) \neq o$, atunci

$$\varphi_r(h) = \frac{h^{s+1}}{(s+1)!}\,\varphi_r^{(s+1)}(\xi) \quad \text{cu} \quad \in (o,h),$$

şi în consecinţă, eroarea comisă în calculul lui y_1 prin formula $y_1 = y_o + A_o(h)$ este

$$e(y_1) \leq \frac{h^{s+1}}{(s+1)!}\,M_{s+1}^r$$

unde $M_{s+1}^r = \sup\left\{ |\varphi_r^{(s+1)}(\xi)|\ ;\ \xi \in ,h] \right\}$.

Funcţia A_o dată de (4.3) corespunde metodei Runge-Kutta de ordinul r. Avînd în vedere că volumul de calcule creşte foarte mult odată cu creşterea ordinului r, cele mai utilizate metode Runge-Kutta sînt cele de ordin $r \leq 4$.

Vom prezenta în continuare cum se determină funcţia A_o pentru cazurile r=1, r=2, r=3 şi r=4.

<u>Metoda Runge-Kutta de ordinul întîi (r=1)</u>

In acest caz (4.3) se rescrie sub forma

$$A_o(h) = p_{11}\,\mathcal{G}_1(h)$$

cu

$$\mathcal{G}_1(h) = h\cdot f(x_o,y_o)\ .$$

Fie $\varphi_1(h) = y(x_o+h) - y(x_o) - p_{11}\,h\cdot f(x_o,y_o)$.

Pentru determinarea constantei p_{11} vom pune condiţia ca derivatele lui φ_1 pînă la un ordin cît mai mare să fie nule în h=o.

Avem

$$\varphi_1'(h) = y'(x_o+h) - p_{11}f(x_o,y_o)$$

şi atunci

$$\varphi_1'(o) = y'(x_o) - p_{11}f(x_o,y_o) = (1-p_{11})f(x_o,y_o) = o$$

dacă şi numai dacă $p_{11}=1$ sau $f(x_o,y_o)=o$.

Pentru a nu restrînge clasa de funcţii f, vom presupune că $f(x_o,y_o) \neq o$ şi atunci $p_{11}=1$. In consecinţă avem

$$A_o(h) = h\cdot f(x_o,y_o)\ .$$

Deci

$$y_1 = y_o + h\cdot f(x_o,y_o),$$

sau, în general

$$y_{i+1} = y_i + h\,f(x_i,y_i),\quad i = \overline{o,n-1}$$

Am obţinut astfel algoritmul de calcul aproximativ al valorilor $y_1,y_2,\ldots,y_n$ propus de metoda lui Euler.

Metoda Runge Kutta de ordinul al doilea (r=2)

In acest caz

$$A_o(h) = p_{21}\,\sigma_1(h) + p_{22}\,\sigma_2(h)$$

unde

$$\sigma_1(h)=h\cdot f(x_o,y_o),\ \sigma_2(h)=h\cdot f(x_o+\alpha_2 h, y_o+\beta_{21}\sigma_1(h)),$$

iar funcţia erorii este

$$\varphi_2(h) = y(x_o+h)-y(x_o)-p_{21}\sigma_1(h)-p_{22}\sigma_2(h)\ .$$

Pentru calculul derivatelor φ_2', φ_2'', φ_2''' ale funcţiei erorii să observăm că

$$\sigma_1'(h) = f(x_o,y_o)= f_o,\quad \sigma_1^{(k)}(h)= o,\ \text{pentru } k \geq 2.$$

iar

$$\sigma_2'(h) = \left[f + h\left(\frac{\partial f}{\partial x}\alpha_2+\frac{\partial f}{\partial y}\beta_{21}\sigma_1'(h)\right)\right]_o$$

$$\sigma_2''(h)= 2\left[\frac{\partial f}{\partial x}\alpha_2+\frac{\partial f}{\partial y}\beta_{21}\sigma_1'(h)\right]_o+h\left[\frac{\partial^2 f}{\partial x^2}\alpha_2^2+\right.$$

$$\left.+\ 2\,\frac{\partial^2 f}{\partial x\partial y}\alpha_2\beta_{21}\sigma_1'(h)+\frac{\partial^2 f}{\partial y^2}\beta_{21}^2\sigma_1'^2(h)\right]_o$$

$$\varphi_2'''(h)=3\left[\frac{\partial^2 f}{\partial x^2}\alpha_2^2 + 2\,\frac{\partial^2 f}{\partial x\partial y}\alpha_2\beta_{21}\sigma_1'(h)+\frac{\partial^2 f}{\partial y^2}\beta_{21}^2\sigma_1'^2(h)\right]_o+$$

$$+\ h.E(h)\ ,$$

unde $[\ \ ]_o$ semnifică faptul că funcţia din paranteză este calculată în $(x,y)=(x_o+\alpha_2 h,\ y_o+\beta_{21}\sigma_1(h))$, iar E este o funcţie a cărei expresie analitică nu ne interesează deoarece în derivatele de mai sus vom lua $h = o$.

Impunînd condiţiile ca $\varphi_2'(o) = \varphi_2''(o)= o$, obţinem

$$\varphi_2'(o)= y'(x_o)-p_{21}f(x_o,y_o)-p_{22}f(x_o,y_o)=(1-p_{21}-p_{22})f(x_o,y_o) = o$$

$$\varphi_2''(0)= \frac{\partial f}{\partial x}(x_0,y_0)+\frac{\partial f}{\partial y}(x_0,y_0)f(x_0,y_0)-2p_{22}\Big[\frac{\partial f}{\partial x}(x_0,y_0)\alpha_2 +$$

$$+\frac{\partial f}{\partial y}(x_0,y_0)f(x_0,y_0)\beta_{21}\Big]= (1-2p_{22}\alpha_2)\frac{\partial f}{\partial x}(x_0,y_0) +$$

$$+ (1-2p_{22}\beta_{21})\frac{\partial f}{\partial y}(x_0,y_0)f(x_0,y_0) = 0 .$$

Pentru a nu restrînge clasa funcţiilor f, vom presupune că $f(x_0,y_0) \neq 0$, $\frac{\partial f}{\partial x}(x_0,y_0) \neq 0$ şi $\frac{\partial f}{\partial y}(x_0,y_0) \neq 0$. In consecinţă

$$\begin{cases} 1 - p_{21} - p_{22} = 0 \\ 1 - 2p_{22}\alpha_2 = 0 \\ 1 - 2p_{22}\beta_{21} = 0. \end{cases}$$

Evident sistemul de mai sus este nedeterminat. De obicei se consideră soluţia $p_{21} = p_{22} = \frac{1}{2}$, $\alpha_2 = \beta_{21} = 1$. Obţinem atunci

$$y_1 = y_0 + \frac{h}{2}\Big[f(x_0,y_0)+f(x_0+h,y_0+ hf(x_0,y_0))\Big],$$

sau, în general

(4.6) $\quad y_{i+1}= y_i+ \frac{h}{2}\Big[f(x_i,y_i)+ f(x_i+h,y_i+h\cdot f(x_i,y_i))\Big], \quad i=\overline{0,n-1}$.

Algoritmul de calcul aproximativ al valorilor $y_1,y_2,\dots,y_n$ dat de (4.6) poartă numele de algoritmul Euler-Cauchy.

<u>Metoda Runge-Kutta de ordinul al treilea (r = 3)</u>

Procedînd ca şi în cazul r=2, se obţine

$$\Delta_0(h) = \frac{1}{6}\Big[\sigma_1(h)+ 4\sigma_2(h)+ \sigma_3(h)\Big],$$

unde

$$\sigma_1(h) = h\cdot f(x_0,y_0),\ \sigma_2(h)= h\cdot f(x_0+ \frac{h}{2},\ y_0+ \frac{\sigma_1(h)}{2}),$$

$$\sigma_3(h) = h\cdot f(x_0+ h,\ y_0+2\sigma_2(h)- \sigma_1(h)).$$

In acest caz algoritmul de calcul aproximativ al valorilor $y_1,y_2,\dots,y_n$ este dat de

$$y_{i+1}=y_i+\frac{h}{6}\cdot f(x_i,y_i)+\frac{2}{3}\,h\cdot f(x_i+\frac{h}{2},y_i+\frac{h}{2}\,f(x_i,y_i))\;+ \tag{4.7}$$

$$+\frac{h}{6}\,f(x_i+h,y_i+2hf(x_i+\frac{h}{2},y_i+\frac{h}{2}\,f(x_i,y_i))-hf(x_i,y_i)),$$

pentru $i=\overline{o,n-1}$.

Metoda Runge-Kutta de ordinul al patrulea (r=4)

In acest caz

$$A_o(h)=\frac{1}{6}\left[\sigma_1(h)+2\,\sigma_2(h)+2\,\sigma_3(h)+\sigma_4(h)\right],$$

unde

$$\sigma_1(h)=hf(x_o,y_o),\ \sigma_2(h)=h\cdot f(x_o+\frac{h}{2},\ y_o+\frac{\sigma_1(h)}{2}),$$

$$\sigma_3(h)=hf(x_o+\frac{h}{2},y_o+\frac{\sigma_2(h)}{2}),\ \ \sigma_4(h)=hf(x_o+h,y_o+\sigma_3(h)).$$

Algoritmul de calcul al valorilor aproximative $y_1,y_2,\dots,y_n$ este dat de

$$y_{i+1}=y_i+\frac{1}{6}\left[\sigma_1(h)+2\,\sigma_2(h)+2\,\sigma_3(h)+\sigma_4(h)\right]\ .$$

$$\sigma_1(h)=h\cdot f(x_i,y_i),\ \ \sigma_2(h)=h\cdot f(x_i+\frac{h}{2},\ y_i+\frac{\sigma_1(h)}{2}), \tag{4.8}$$

$$\sigma_3(h)=h\ f(x_i+\frac{h}{2},\ y_i+\frac{\sigma_2(h)}{2}),\ \ \sigma_4(h)=h\ f(x_i+h,y_i+\sigma_3(h)).$$

Incheiem această secţiune cu observaţia că şi în cazul metodei Runge-Kutta, evaluarea erorii se poate face cu ajutorul Teoremei 2.1.

§ 5. Metode cu diferenţe finite. Metoda extrapolării a lui Adams

Să considerăm problema lui Cauchy

$$\begin{cases}\dfrac{dy}{dx}\,(x)=f(x,y(x)), & x_o\le x\le x_o+\delta\\ y(x_o)=y_o, & \end{cases} \tag{5.1}$$

unde $f:[x_0,x_0+\delta]\times R\to R$ este o funcţie de clasă C^k (k fiind suficient de mare), iar $y_0\in R$.

Să considerăm diviziunea (P): $x_0 < x_1 < \ldots < x_n = x_0+\mu$ a intervalului $[x_0,x_0+\mu]$ de existenţă a soluţiei y. In continuare vom presupune că $x_{i+1}-x_i = h$ - constant - pentru $i=\overline{0,n-1}$.

Să presupunem că am determinat valorile aproximative $y_i = y(x_i)$ pentru $i = r-p, r-p+1,\ldots,r-1,r$ $(0<p<r)$ printr-una din metodele descrise anterior şi că intenţionăm să determinăm valoarea $y_{r+1} = y(x_{r+1})$.

In acest scop, să observăm că soluţia „exactă" y a problemei (5.1) satisface

(5.2) $$y(x_{r+1}) = y(x_r)+\int_{x_r}^{x_{r+1}} f(x,y(x))dx .$$

Inlocuind $f(x,y(x))$ prin polinomul de interpolare Newton cu diferenţe retrograde (vezi cap.V, § 4, formula (4.10)):

$$f(x,y(x))= f_r+\frac{t}{1!}\nabla f_r+\frac{t(t+1)}{2!}\nabla^2 f_r+\ldots+\frac{t(t+1)\ldots(t+p-1)}{p!}\nabla^p f_r$$

unde $t=\dfrac{x-x_r}{h}$ şi $f_r = f(x_r,y_r)$, obţinem

(5.3) $$y_{r+1}= y_r + h\sum_{q=0}^{p}\beta_q\nabla^q f_r$$

Menţionăm că relaţia (5.3) β_q sînt daţi de

$$\beta_q = \frac{1}{q!}\int_0^1 t(t+1)\ldots(t+q-1)dt\ ,\quad q = 1,2,\ldots,\ \beta_0 = 1.$$

Eroarea comisă prin calculul lui y_{r+1} cu ajutorul formulei (5.3) este (pentru h suficient de mic !)

(5.4) $$e(y_{r+1})\le h^{p+2}\beta_{p+1}M_{p+1}\ ,$$

unde $M_{p+1} = \sup\{|f^{(p+1)}(x,z(x))|;\ x\in[x_{r-p},x_r]\}$, z fiind soluţia aproximativă a problemei (5.1) cunoscută deja pe $[x_{r-p},x_r]$.

Intrucît $\beta_0 = 1$, $\beta_1 = \frac{1}{2}$, $\beta_2 = \frac{5}{12}$, $\beta_3 = \frac{3}{8}$, $\beta_4 = \frac{251}{720}$, $\beta_5 = \frac{19087}{60480}$ etc., (5.3) se poate scrie sub forma

$$y_{r+1} = y_r + h\left[f_r + \frac{1}{2}\nabla f_r + \frac{5}{12}\nabla^2 f_r + \frac{3}{8}\nabla^3 f_r + \frac{251}{729}\nabla^4 f_r + \ldots\right].$$

Dacă în (5.3) exprimăm $\nabla^q f_r$ în funcţie de valorile $f_{r-p}, f_{r-p+1}, \ldots, f_r$, obţinem aşa numita formă a lui Lagrange a relaţiei (5.3):

$$y_{r+1} = y_r + h\sum_{q=0}^{p} \alpha_{pq} f_{r-q} \tag{5.5}$$

Menţionăm că pentru p dat α_{pq} se determină prin identificarea coeficienţilor. Cîteva valori pentru α_{pq} sînt date de

α_{pq}

p	q = o	1	2	3
1	$\frac{3}{2}$	$-\frac{1}{2}$	-	-
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$	-
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{34}{24}$	$-\frac{9}{24}$

§ 6. Metoda interpolativă a lui Adams

Această metodă diferă de precedenta prin aceea că aici presupunem cunoscute valorile $f_{r-p+1}, f_{r-p+2}, \ldots, f_r, f_{r+1}$ ale funcţiei $f(x,y(x))$ calculate în $x_{1-p+1}, x_{r-p+2}, \ldots, x_r, x_{r+1}$.

Inlocuind $f(x,y(x))$ în (5.2) prin polinomul de interpolare Newton cu diferenţe regresive (vezi Cap.V, § 4, formula (4.10)):

$$f(x,y(x)) = f_{r+1} + \frac{t-1}{1!}\nabla f_{r+1} + \frac{(t-1)t}{2!}\nabla^2 f_{r+1} + \ldots +$$
$$+ \frac{(t-1)t(t+1)\ldots(t+p-2)}{p!}\nabla^p f_{r+1}.$$

unde $t = \frac{x-x_r}{h}$, obţinem

$$(6.1)\qquad y(x_{r+1}) = y(x_r) + h\sum_{q=0}^{p} \beta_q^* \nabla^q f(x_{r+1}, y(x_{r+1})) ,$$

unde

$$\beta_q^* = \frac{1}{q!}\int_0^1 (t-1)t(t+1)\dots(t+q-2)dt , \quad \beta_o^* = 1 .$$

Cu notaţiile $y(x_{r+1}) = y_{r+1}$, $y(x_r) = y_r$, $f(x_{r+1}, y_{r+1}) = f_{r+1}$, formula (6.1) se poate rescrie - înlocuind valorile coeficienţilor β_q^* - sub forma

$$y_{r+1} = y_r + h\left[f_{r+1} - \frac{1}{2}\nabla f_{r+1} - \frac{1}{12}\nabla^2 f_{r+1} - \frac{1}{24}\nabla^3 f_{r+1}\dots\right]$$

Subliniem că dificultatea majoră în calculul lui y_{r+1} cu ajutorul formulei (6.1) constă în aceea că, spre deosebire de (5.3) care ne furnizează valoarea <u>explicită</u>, a lui y_{r+1}, (6.1) defineşte y_{r+1} <u>în mod implicit</u>. Cu toate acestea, pentru h suficient de mic ecuaţia neliniară (6.1) cu necunoscuta y_{r+1} se poate rezolva prin metoda aproximaţiilor succesive după cum urmează. Pornind de la o valoare aproximativ $y_{r+1}^{[s]}$ în membrul drept calculăm $f_{r+1}^{[s]} = f(x_{r+1}, y_{r+1}^{[s]})$ şi $\nabla^q f_{r+1}^{[s]} = \nabla^{q-1} f_{r+1}^{[s]} - \nabla^{q-1} f_r^{[s]}$ şi apoi determinăm

$$(6.2)\qquad y_{r+1}^{[s+1]} = y_r + h\sum_{q=0}^{p} \beta_q^* \nabla^q f_{r+1}^{[s]} , \qquad s = 0,1,2,\dots .$$

Ca şi în cazul formulei (5.3), (6.1) este uneori utilizată sub aşa numita formă a lui Lagrange:

$$(6.3)\qquad y_{r+1} = y_r + h\sum_{q=0}^{p} \alpha_{pq}^* f_{r+1-q}$$

unde $\alpha_{po}^* = \beta_p$, iar α_{pq}^* se calculează prin identificare. Cîteva valori ale coeficienţilor α_{pq}^* sînt date de tabelul de mai jos

α^{*}_{pq}

p	q = o	1	2	3	4
1	$\frac{1}{2}$	$\frac{1}{2}$	-	-	-
2	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$	-	-
3	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	-
4	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

Eroarea comisă prin calculul lui v_{r+1} cu ajutorul formulei (6.1) este (pentru h suficient de mic!)

$$(6.4) \qquad e(v_{r+1}) \leq h^{p+2} \; |\beta^{*}_{p+1}| M_{p+1}$$

unde $M_{p+1} = \sup\left\{ |f^{(p+1)}(x,z(x))| ;\; x \in [x_{r-p+1}, x_{r+1}] \right\}$, z fiind soluţia aproximativă a problemei (5.1) pe $[x_{r-p+1}, x_{r+1}]$.

Facem observaţia că atît pentru metoda extrapolativă a lui Adams cît şi pentru metoda interpolativă a lui Adams, evaluarea erorii se poate face cu ajutorul Teoremei 2.1.

Incheiem această secţiune cu următoarele „reguli" de utilizare ale diverselor metode, „reguli" impuse de practică.

(A). Metodele cu diferenţe finite se dovedesc a fi foarte potrivite atunci cînd f este suficient de regulată şi cînd diferenţele finite descresc rapid odată cu creşterea ordinului. In această situaţie, chiar pentru un pas h suficient de mic, toate calculele implicate la folosirea formulelor (5.3), (6.1), (6.2) se pot efectua cu destulă uşurinţă.

(B). Dacă f nu este suficient de regulată, sau dacă dorim să folosim un pas h nu prea mic, atunci metodele de tip Runge - Kutta sînt de preferat. De asemenea, este avantajos să utilizăm metodele de tip Runge-Kutta ori de cîte ori dorim să modificăm pasul h, mai ales atunci cînd această modificare revine la o micşorare.

(C). O mărime semnificativă care măsoară acuratețea aproximării este

$$S = k.h$$

unde $k = \left|\frac{\partial f}{\partial y}(x_i, y_i)\right|$, iar h este pasul diviziunii.

Pentru obținerea unei aproximări cu un grad moderat de acuratețe este necesar ca S să fie între 0,05 și 0,1. In consecință, dacă după un anumit număr de pași S depășește 0,1, este indicat să micșorăm pasul.

Pentru alte detalii a se consulta L.Collatz [3, Ch.II, § 3, p. 78-97].

§ 7. Metoda lui Ritz de rezolvare numerică a problemelor la limită pentru ecuații diferențiale de ordinul al doilea

In acest paragraf ne propunem să descriem o metodă de rezolvare numerică a problemelor la limită de tipul

$$L[y] = -\frac{d}{dx}[p(x)y'(x)] + q(x)y(x) = r(x), \quad a \leq x \leq b$$

(7.1) $$U_1[y] = \alpha_0 y(a) + \alpha_1 y'(a) = \gamma_1$$

$$U_2[y] = \beta_0 y(b) + \beta_1 y'(b) = \gamma_2 ,$$

unde p,q,r sînt funcții de clasă C^2 pe [a,b], iar $\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma_1$ și γ_2 sînt constante date.

Se știe că (7.1) reprezintă ecuația Euler-Lagrange pentru problema de minim

(7.2) $$\min\left\{ \int_a^b (p\varphi'^2 + q\varphi^2 - 2r\varphi)dx + A_0\varphi^2(a) + B_0\varphi^2(b) + 2A_1\varphi(a) + 2B_1\varphi(b) ;\ \varphi \in \mathcal{F} \right\}$$

unde A_0, B_0, A_1, $B_1 \in R$ și $F \subset C^2[a,b]$ sînt alese după cum urmează.

Cazul I. Dacă $\alpha_1 \neq o$ şi $\beta_1 \neq o$, atunci

(7.3) $A_o = -\frac{\alpha_o}{\alpha_1} p(a)$, $A_1 = \frac{\gamma_1}{\alpha_1} p(a)$, $B_o = \frac{\beta_o}{\beta_1} p(b)$, $B_1 = -\frac{\gamma_2}{\beta_1} p(b)$

şi $F = C^2[a,b]$.

Cazul II. Dacă $\alpha_1 = o$, $\beta_1 = o$, dar $\alpha_o \neq o$, atunci

(7.4) $A_o = A_1 = B_o = B_1 = o$ şi $F = \{\varphi \in C^2[a,b];\ \varphi(a) = \frac{\gamma_1}{\alpha_o},$
$\varphi(b) = \frac{\gamma_2}{\beta_o}\}$.

Cazul III. Dacă $\alpha_1 = o$ şi $\beta_1 \neq o$, atunci

(7.5) $A_o = A_1 = o$, $B_o = \frac{\beta_o}{\beta_1} p(b)$, $B_1 = -\frac{\gamma_2}{\beta_1} p(b)$ şi $F = \{\varphi \in C^2[a,b];$
$\alpha_o \varphi(a) = \gamma_1\}$.

Similar, dacă $\alpha_1 \neq o$ şi $\beta_1 = o$, atunci

(7.6) $A_o = -\frac{\alpha_o}{\alpha_1} p(a)$, $A_1 = \frac{\gamma_1}{\alpha_1} p(a)$, $B_o = B_1 = o$ şi $F = \{\varphi \in C^2[a,b];$
$\beta_o \varphi(b) = \gamma_2\}$.

A se vedea L.Collatz [3, p.206].

Convenim ca, în cele ce urmează, să numim condiţii la limită esenţiale condiţiile impuse funcţiilor φ în definirea mulţimii F. De exemplu, în Cazul I nu avem condiţii la limită esenţiale, iar în Cazul al II-lea acestea sînt $\varphi(a) = \frac{\gamma_1}{\alpha_o}$, $\varphi(b) = \frac{\gamma_2}{\beta_o}$.

Ideea care stă la baza metodei lui Ritz constă în a înlocui (7.1) cu problema corespunzătoare de minim (7.2) şi apoi în a căuta soluţia aproximativă φ sub forma

(7.7) $$\varphi(x) = v_o(x) + \sum_{n=1}^{k} a_n v_n(x)$$

unde $v_o, v_1, \ldots, v_n$ sînt funcţii date, de clasă C^2, alese astfel încît dacă în alegerea lui F avem condiţie la limită esenţială,

v_o să satisfacă această condiţie la limită (în a,b, sau în a şi în b), iar v_n să satisfacă $v_n(a) = o$, sau $v_n(b)= o$, sau $v_n(a)= = v_n(b) = o$ - după cum condiţia la limită esenţială este în a,b sau şi în a şi în b.

In consecinţă, problema se reduce la determinarea coeficienţilor $a_1,a_2,\dots,a_n$ astfel încît funcţia corespunzătoare dată de (7.7) să realizeze minimul din (7.2) calculat <u>numai pentru funcţii φ de forma (7.7)</u>.

Altfel spus, ne interesează să găsim φ^* care satisface

$$J(\varphi^*) = \min\{J(\varphi);\ \varphi \text{ de forma } (7,7)\} \tag{7.8}$$

unde

$$J(\varphi)= \int_a^b(p\varphi'^2+ q\varphi^2- 2r\varphi)dx+ A_o\varphi^2(a)+B_o\varphi^2(b) + $$
$$+ 2A_1\varphi(a)+ 2B_1\varphi(b).$$

In consecinţă, coeficienţii $a_1,a_2,\dots,a_k$ ai lui φ^* de forma (7.7) trebuie să satisfacă

$$\frac{1}{2}\,\frac{\partial J(\varphi^*)}{\partial a_i} = o\ , \quad i = \overline{1,k}\ ,$$

sistem care poate fi rescris sub forma

$$\int_a^b(p\varphi^{*\prime}v_i' + q\varphi^* v_i- rv_i)dx + \frac{p(a)v_i(a)}{\alpha_1}\left[-\alpha_o\varphi^*(a)+\gamma_1\right]+ \tag{7.9}$$
$$+\frac{p(b)v_i(b)}{\beta_1}\left[\beta_o\varphi^*(b)-\gamma_2\right]= o, \quad i = \overline{1,k}\ .$$

Integrînd prin părţi obţinem

$$\int_a^b v_i\{ L[\varphi^*] -r\}dx - \frac{p(a)v_i(a)}{\alpha_1}\{U_1[\varphi^*]- \gamma_1\} + \tag{7.10}$$
$$+\frac{p(b)v_i(b)}{\beta_1}\{U_2[\varphi^*] - \gamma_2\}= o, \quad i = \overline{1,k}\ ,$$

unde L, U_1 şi U_2 sînt definiţi ca în (7.1).

Inlocuind $\varphi^* = v_o + \sum_{j=1}^{k} a_j v_j$ în (7.10) deducem

$$(7.11)\qquad \sum_{j=1}^{k} a_j \left\{ \int_a^b v_i L[v_j]dx - \frac{p(a)v_i(a)}{\alpha_1} U_1[v_j] + \frac{p(b)v_i(b)}{\beta_1} U_2[v_j] \right\} =$$

$$= - \int_a^b v_i\{L[v_o]-r\}dx + \frac{p(a)v_i(a)}{\alpha_1}\{U_1[v_o]-\gamma_1\} -$$

$$- \frac{p(b)v_i(b)}{\beta_1}\{U_2[v_o]-\gamma_2\}, \qquad i = \overline{1,k}\ ,$$

sistem din care se determină $a_1, a_2, \ldots, a_k$.

Incheiem acest paragraf printr-un exemplu ilustrativ.

<u>Exemplul 7.1.</u> Să considerăm problema - vezi L.Collatz [3 , p.209].

$$y''(x) + (1+x^2)y(x)+1 = 0\ , \qquad -1 \leq x \leq 1$$

$$y(\pm 1) = 0.$$

In acest caz avem $a = -1$, $b = 1$, $p = 1$, $q(x) = -(1+x^2)$, $r=1$, $\alpha_o = 1$, $\alpha_1 = 0$, $\gamma_1 = 0$, $\beta_o = 1$, $\beta_1 = 0$, $\gamma_2 = 0$ şi atunci

$$J(\varphi) = \int_{-1}^{1} [\varphi'^2 - (1+x^2)\varphi^2 - 2\varphi]\ dx\ ,$$

iar $F = \{\varphi \in C^2[-1,1];\ \varphi(-1) = \varphi(1) = 0\}$. Vezi (7.4).

Pentru simplitate, vom considera $k=2$, $v_o = 0$, $v_1(x) = 1-x^2$ şi $v_2(x) = 1-x^4$ pentru $x \in [-1,1]$, care evident satisfac $v_1(\pm 1) = v_2(\pm 1) = 0$.

Deci

$$\varphi^*(x) = a_1(1-x^2) + a_2(1-x^4),$$

şi

$$\frac{1}{8} J(\varphi^*) = \frac{19}{105} a_1^2 + \frac{10}{45} 2a_1a_2 + \frac{1244}{3465} a_2^2 - \frac{1}{3} a_1 - \frac{2}{5} a_2\ .$$

In consecinţă,

$$\frac{1}{8} \frac{\partial J(\varphi^*)}{\partial a_1} = \frac{38}{105} a_1 + \frac{20}{45} a_2 - \frac{1}{3} = 0$$

$$\frac{1}{8} \frac{\partial J(\varphi^*)}{\partial a_2} = \frac{20}{45} a_1 + \frac{2488}{3465} a_2 - \frac{2}{5} = 0\ ,$$

de unde $a_1 = 0,98777$ şi $a_2 = -0,05433$. Deci soluţia aproximativă a problemei este

$$v(x) = \varphi^*(x) = 0,98777\,(1-x^2) - 0,05433\,(1-x^4),$$

pentru $x \in [-1,1]$.

§ 8. Metode cu diferenţe finite pentru rezolvarea ecuaţiilor cu derivate parţiale

Să considerăm ecuaţia cu derivate parţiale de ordinul al doilea, liniară

$$(8.1)\quad a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} + fu + g = 0, \quad (x,y) \in D,$$

unde a,b,c,d,e,f şi g sînt funcţii continue pe domeniul mărginit $D \subset R^2$.

Intrucît nu întotdeauna se poate determina în mod explicit o soluţie a ecuaţiei (8.1) care să satisfacă anumite condiţii suplimentare: condiţii iniţiale, condiţii limită, este necesar să se pună în evidenţă o metodă de calcul aproximativ al valorilor soluţiei într-o reţea de puncte din domeniul $D \subset R^2$.

Descriem în continuare o astfel de metodă bazată pe înlocuirea derivatelor parţiale din (8.1) cu diferenţe divizate.

Incepem cu cazul eliptic. Să considerăm problema

$$(8.2)\quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y) \quad \text{pentru} \quad (x,y) \in D$$

$$u(x,y) = g(x,y) \quad \text{pentru} \quad (x,y) \in C \text{ - frontiera lui } D,$$

unde f şi g sînt funcţii continue pe D, şi respectiv C.

Să împărţim R^2 printr-o reţea de drepte paralele cu axele într-o mulţime de dreptunghiuri dintre care le reţinem numai pe acelea care au intersecţia nevidă cu $\bar{D} = D \cup C$. (vezi figura 8.1).

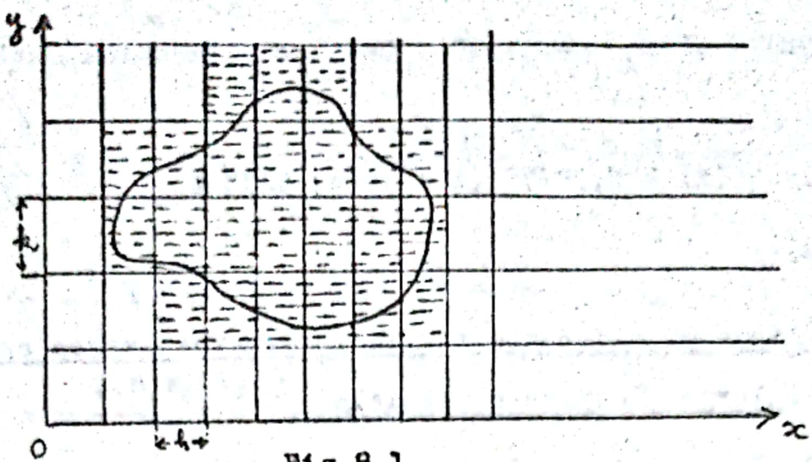


Fig.8.1.

Pentru simplitate, în cele ce urmează, vom considera h = k. Nodurile (i,j) ale rețelei care aparțin lui D le împărțim în două clase.

1. Clasa nodurilor (i,j) din D pentru care toate cele patru noduri imediat vecine (i-1,j), (i+1,j), (i,j-1) și (i,j+1) sînt din $\overline{D}$. Să notăm cu n_1 numărul acestor noduri.

2. Clasele acelor noduri (i,j) din D pentru care cel puțin unul din cele patru noduri imediat vecine nu este în $\overline{D}$. Să notăm cu n_2 numărul acestor noduri.

In continuare vom arăta cum pot fi calculate valorile aproximative ale soluției u în nodurile rețelei care aparțin lui $\overline{D}$.

Să notăm cu $u_{ij}= u(ih,jh)$, cu $f_{ij}= f(ih,jh)$ și cu $g_{ij}=g(ih,jh)$

Dacă (i,j) este un nod din prima clasă, înlocuim ecuația (8.2) prin

(8.3) $$\frac{1}{h^2}(u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}-4u_{ij}) = f_{ij} .$$

Dacă (i,j) este un nod din cea de a doua clasă, atunci înlocuim nodurile imediat vecine care nu sînt în D prin punctele de intersecție ai rețelei cu frontiera C, cele mai apropiate de (i,j). De pildă, dacă $M_0(x_0,y_0)$ este un nod din cea de a doua clasă, atunci cele patru puncte imediat vecine pe care le vom lua în discuție vor fi $M_1(x_0+t_1h,y_0)$, $M_2(x_0-t_2h,y_0)$, $M_3(x_0,y_0+t_3h)$ și $M_4(x_0,y_0-t_4h)$ cu $0<t_i\le 1$, $i=\overline{1,4}$. Folosind dezvoltarea în

serie Taylor a lui u în jurul punctelor M_i, $i=\overline{1,4}$ şi neglijînd resturile (de ordinul lui h^3) se obţine

$$(8.4)\quad \frac{1}{t_1+t_2}\left(\frac{u_1}{t_1}+\frac{u_2}{t_2}\right)+\frac{1}{t_3+t_4}\left(\frac{u_3}{t_3}+\frac{u_4}{t_4}\right)-\left(\frac{1}{t_1+t_2}+\frac{1}{t_3+t_4}\right)u_o = \left(\frac{h^2}{2}\right) f(x_o,y_o),$$

unde u_i este valoarea lui u în M_i, $i=\overline{1,4}$.

In acest fel vom obţine prin cuplarea lui (8.3) cu (8.4) un sistem de $n = n_1+n_2$ ecuaţii liniare cu n necunoscute - u_{ij} - valorile aproximative ale lui u în nodurile reţelei.

Subliniem că în cazul considerat de noi acest sistem are soluţie unică.

Să considerăm acum ecuaţia parabolică

$$(8.5)\quad \begin{cases} \dfrac{\partial u}{\partial t} = a^2\dfrac{\partial^2 u}{\partial x^2}, & o\le x\le 1, \quad o\le t\le 1 \\ u(o,t) = u(1,t) = o, & o\le t\le 1 \\ u(x,o) = f(x), & o\le x\le 1, \end{cases}$$

unde f este o funcţie continuă cu $f(o) = f(1) = o$.

Reamintîm că (8.5) reprezintă modelul matematic ce descrie propagarea căldurii într-o bară omogenă de lungime 1, cu temperatura zero la capete; $u(x,t)$ fiind temperatura secţiunii de abscisă x a barei la momentul t.

In acest caz $\overline{D} = [o,1]\times[o,1]$ poate fi împărţit printr-o reţea de drepte paralele cu axele astfel încît nodurile din cea de a doua clasă să se găsească pe frontiera C a lui D (vezi fig.8.2).

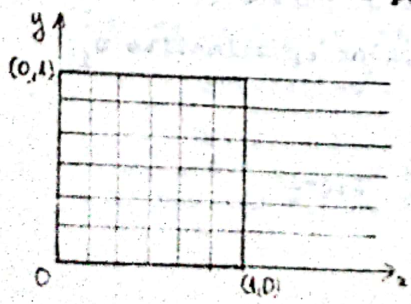


Fig.8.2.

Vom descrie mai departe un algoritm de rezolvare aproximativă a ecuației (8.5) care pentru h şi k tinzînd la o este convergent. Intelegem prin aceasta că dacă $h \to o$, $k \to o$, $(ih, jk) \to (x,y) \in \overline{D}$, atunci $u_{ij} \to u(x,y)$ - valoarea exactă a soluției u în (x,y).

Acest algoritm de calcul al valorilor aproximative u_{ij} este dat de sistemul

$$(8.6) \qquad (2+2r)u_{i,j+1} - ru_{i-1,j+1} - ru_{i+1,j+1} = ru_{i-1,j} + (2-2r)u_{ij} + ru_{i+1,j} \ ,$$

unde $r = \frac{a^2k}{h^2}$. (Algoritmul Crank-Nicolson).

Pentru alți algoritmi de rezolvare aproximativă a problemei (8.5) vezi B.Crstici şi alții [5, p.425-434].

In sfîrşit, să considerăm ecuația hiperbolică

$$(8.7) \qquad \begin{cases} \dfrac{\partial^2 u}{\partial t^2} = a^2 \dfrac{\partial^2 u}{\partial x^2} \ , & o \le x \le 1 \ , \quad o \le t \le 1 \\ u(o,t) = u(1,t) = o \ , & o \le t \le 1 \\ u(x,o) = f(x) \ , & o \le x \le 1 \\ \dfrac{\partial u}{\partial t}(x,o) = o \ , & o \le x \le 1 \ , \end{cases}$$

unde f este o funcție continuă cu $f(o) = f(1) = o$.

Reamintim că (8.7) descrie vibrațiile unei corzi de lungime 1 fixată la capete; $u(x,t)$ fiind elongația punctului de abscisă x de pe coardă la momentul t, iar $f(x)$ este elongația aceluiaşi punct de pe coardă la $t = o$.

Folosind aceeaşi rețea ca şi în cazul parabolic - vezi figura 8.2 - algoritmul de calcul a valorilor aproximative u_{ij} este dat de

$$(8.8) \qquad u_{i+1,j} - 2u_{i,j} + u_{i-1,j} - \frac{1}{r}\left(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}\right) = o \ ,$$

unde $r = \frac{a^2k^2}{h^2}$.

Luînd $h = \frac{1}{n}$, alegînd k astfel încît $r = 1$ şi scoţînd $u_{i,j+1}$ din (8,8), obţinem

(8.9) $$u_{i,j+1} = (u_{i+1,j} + u_{i-1,j}) - u_{i,j-1}.$$

Condiţiile la limită şi condiţiile iniţiale se scriu sub forma

(8.10) $$u_{o,j} = u_{n,j} = o, \quad u_{i,o} = f_i, \; u_{i,1} = u_{i,-1} \quad (\text{cu } f_i = f(h_i)).$$

Luînd $j = o$ şi folosind (8.10) obţinem din (8,9)

(8.11) $$u_{i,1} = \frac{1}{2}(u_{i+1,o} + u_{i-1,o}) = \frac{1}{2}(f_{i+1} + f_{i-1}).$$

Din (8.9), (8.10) şi (8.11) putem afla toate valorile aproximative $u_{i,j}$.

§ 9. Exerciţii

1. Să se scrie schema logică pentru rezolvarea ecuaţiei diferenţiale $y'(x) = f(x,y(x))$, $x_o \leq x \leq x_o + \mu$, $y(x_o) = y_o$ cu ajutorul metodei lui Euler.

2. Aceeaşi problemă pentru cazul metodelor Runge-Kutta de ordinul 2, 3 şi 4 respectiv.

3. Aceeaşi problemă pentru metoda extrapolativă a lui Adams în cazul $p = 3$. Primele 4 valori se vor calcula cu ajutorul metodei Runge-Kutta de ordinul 4.

4. Aceeaşi problemă pentru metoda interpolativă a lui Adams în cazul $p = 3$. Primele patru valori se vor calcula cu ajutorul metodei Runge-Kutta de ordinul 4.

5. Să se rezolve cu ajutorul metodei lui Ritz problema

$$y''(x) + (1-x^2)y(x) + 2y(x) = o, \quad -1 \leq x \leq 1$$
$$y(-1) = y(1) = o$$

Se vor lua $v_o = o$, $v_1(x) = 1-x^2$ şi $v_2(x) = 1-x^4$.

6. Să se scrie schema logică de rezolvare a problemei

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = (1-x^2)(1-y^2), & (x,y) \in (-1,1)x(-1,1) \\ u(-1,y) = u(1,y) = u(x,1) = u(x,-1) = o \end{cases}$$

prin algoritmul (8.3)-(8.4) în cazul $h = k = \frac{1}{5}$.

7. Să se scrie schema logică de rezolvare a problemei

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}, & o \leq x \leq 1, \quad o \leq t \leq 1 \\ u(o,t) = u(1,t) = o, & o \leq t \leq 1 \\ u(x,o) = x(1-x^2), & o \leq x \leq 1 \end{cases}$$

prin algoritmul (8.6) cu $k = h = \frac{1}{10}$.

8. Să se scrie schema logică de rezolvare a problemei

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = \dfrac{\partial^2 u}{\partial x^2}, & o \leq x \leq 1, \quad o \leq t \leq 1 \\ u(o,t) = u(1,t) = o, & o \leq t \leq 1 \\ u(x,o) = x(1-x), & o \leq x \leq 1 \\ \dfrac{\partial u}{\partial t}(x,o) = o, & o \leq x \leq 1 \end{cases}$$

prin algoritmul (8.9), (8.10), (8.11) cu $k = h = \frac{1}{10}$.

BIBLIOGRAFIE

1 Bucur, C.M., Poplea, C.A., Simion, Gh.Gh. - Matematici speciale - Calcul numeric, E.D.P., Bucureşti (1983).

2 Burdeu, R.L., Paires, J.D., Reynolds, A.C. - Numerical Analysis, Boston - Massachusetts, Prindle, Weler & Schmidt (1978).

3 Collatz, L. - The numerical treatment of differential equations, Springer Verlag (1966).

4 Corduneanu, A. - Ecuaţii diferenţiale cu aplicaţii în electrotehnică, Ed. Facla, Timişoara (1981).

5 Crstici, B. şi alţii - Matematici speciale, E.D.P.Bucureşti (1981).

6 Cuculescu, I. - Analiză numerică, Editura Tehnică, Bucureşti, (1967).

7 Demidovich, B.P., Mason, I.A. - Computational mathematics, MIR, Moscou (1973).

8 Dodescu, Gh. - Metode numerice în algebră, Editura Tehnică, Bucureşti (1979).

9 Fröberg, C.-E. - Instruction to numerical analysis, Addison - Wesley (1965).

10 Ilioi C. - Analiză numerică, Partea I, Univ. „Al.I.Cuza", Iaşi, Rotaprint (1982).

11 Ionescu, D.V. - Cuadraturi numerice, Editura Tehnică, Bucureşti (1957).

12 Macon, N. - Numerical Analysis, John Wiley (1962).

13 Marinescu, Gh. - Analiză numerică, Editura Acad. R.S.R.(1974).

14 Marinescu, Gh. şi alţii - Probleme de analiză numerică rezolvate pe calculator, Editura Acad. R.S.R.(1987).

15 Marinescu, Gh. - Analiza matematică, Vol.I, Editura Acad. R.S.R. (1983).

16 Micula, Gh. - Funcţii spline şi aplicaţii. Editura Tehnică, Bucureşti (1978).

17 Murgescu V. - Curs de analiză matematică şi matematici speciale, Vol.III, Capitole de matematici speciale, Rotaprint, I.P.Iaşi (1977).

18 Păvăloiu, I. - Rezolvarea ecuaţiilor prin interpolare, Editura Dacia, Cluj - Napoca (1981).

19 Sabac, I.Gh. - Matematici speciale, Vol.I, E.D.P. Bucureşti (1981).

20 Sabac, I.Gh. şi alţii - Matematici speciale, Vol.II, E.D.P. Bucureşti (1983).

21 Voievodine, V. - Principes numériques d'algebre linéaire, MIR, Coscou (1980).

22 Wilkinson, I.H., - The algebraic eigenvalue problem, Oxford University Press (1965).

C U P R I N S

Pag.

Cap.I. NOTIUNI INTRODUCTIVE

§ 1. Inegalităţi. 5
§ 2. Integrala Lebesgue 10
§ 3. Spaţii metrice. Teorema de punct fix a lui Banach. 27
§ 4. Spaţii liniare normate, Spaţii Banach şi Spaţii Hilbert 37
§ 5. Operatori liniari şi matrici 46
§ 6. Elemente de teoria erorilor. 53
§ 7. Exerciţii. 58

Cap.II. IMBUNATATIREA VITEZEI DE CONVERGENTA A SERIILOR

§ 1. Aproximarea sumei seriilor numerice convergente. 62
§ 2. Imbunătăţirea vitezei de convergenţă a seriilor numerice. 70
§ 3. Imbunătăţirea vitezei de convergenţă a seriilor întregi prin metoda Euler Abel 73
§ 4. Imbunătăţirea vitezei de convergenţă a seriilor Fourier prin metoda lui Krîlov 75
§ 5. Sumarea aproximativă a seriilor trigonometrice. 78
§ 6. Exerciţii. 80

Cap.III. REZOLVAREA NUMERICA A SISTEMELOR DE ECUATII

A. Rezolvarea numerică a sistemelor de ecuaţii algebrice liniare. 81
§ 1. Generalităţi 81
§ 2. Rezolvarea sistemelor de formă specială. 82
§ 3. Factorizarea matricilor. Inversa generalizată a unei matrici. 86
§ 4. Procedeul de tridiagonalizare a lui Lanczos. . . 1o6
§ 5. Metode directe de rezolvare a sistemelor de ecuaţii liniare. 1o9

Pag.

§ 6. Metode iterative de rezolvare a sistemelor de ecuaţii liniare. 114
§ 7. Soluţii generale ale unui sistem de ecuaţii liniare. 130
§ 8. Sisteme bine condiţionate şi rău condiţionate. Măsuri ale condiţionării. 132
§ 9. Exerciţii. 136

B. Rezolvarea numerică a ecuaţiei f(x)= o. 139
§ 1. Generalităţi. 139
§ 2. Localizarea rădăcinilor 140
§ 3. Metode de determinare a rădăcinilor unei ecuaţii algebrice 145
§ 4. Metode iterative de rezolvare a ecuaţiei f(x) = o. 150
§ 5. Exerciţii 158

C. Rezolvarea numerică a sistemelor de ecuaţii neliniare. 160
§ 1. Metoda lui Newton 160
§ 2. Metoda gradientului 164
§ 3. Exerciţii. 168

Cap.IV. METODE NUMERICE DE DETERMINARE A VALORILOR PROPRII
§ 1. Generalităţi. 169
§ 2. Localizarea valorilor proprii 172
§ 3. Metoda lui Danilevski de reducere a unei matrici de forma Frobenius. 176
§ 4. Metoda lui Jacobi pentru matrici reale simetrice. 177
§ 5. Algoritmii LU şi QR. 185
§ 6. Metoda puterilor. 187
§ 7. Exerciţii. 191

Cap.V. INTERPOLARE
§ 1. Formula de interpolare a lui Lagrange . . . 195
§ 2. Formula de interpolare a lui Hermite. . . . 205
§ 3. Diferenţe finite şi diferenţe divizate. . . 211
§ 4. Formula de interpolare a lui Newton.Formule de interpolare pe noduri echidistante . . . 221

Pag.
§ 5. Funcţii spline. 227
§ 6. Exerciţii. 236
Cap. VI. INTEGRARE NUMERICA
§ 1. Generalităţi asupra formulelor de cuadratură. 240
§ 2. Formule de cuadratură Newton-Côtes simple. 244
A. Inchise. 244
B. Deschise 250
§ 3. Forma restului formulelor Newton-Côtes simple. 254
§ 4. Formule de cuadratură Newton-Côtes iterate. 260
§ 5. Formule de cuadratură de tip Gauss. . . . 266
§ 6. Formule de cuadratură Gauss-Legendre, Gauss-Laguerre şi Gauss-Hermite 272
§ 7. Observaţii asupra erorilor şi convergenţei formulelor de cuadratură. 281
§ 8. Exerciţii. 287
Cap.VII. DERIVARE NUMERICA
§ 1. Formule de derivare numerică în noduri arbitrare. 291
§ 2. Formule de derivare numerică în noduri echidistante 295
§ 3. Observaţii asupra influenţei erorilor de rotunjire 301
§ 4. Exerciţii. 303
Cap.VIII. METODE DE REZOLVARE NUMERICA A ECUATIILOR DIFERENTIALE ORDINARE SI CU DERIVATE PARTIALE
§ 1. Noţiuni introductive. 307
§ 2. Soluţii ε-aproximative 312
§ 3. Metoda lui Euler (a liniilor poligonale) de rezolvare numerică a ecuaţiilor diferenţiale de ordinul întîi. 317
§ 4. Metoda Runge-Kutta de rezolvare numerică a ecuaţiilor diferenţiale de ordinul întîi.. 325

Pag.

§ 5. Metode cu diferenţe finite. Metoda extrapolativă a lui Adams. 331
§ 6. Metoda interpolativă a lui Adams. 333
§ 7. Metoda lui Ritz de rezolvare numerică a problemelor la limită pentru ecuaţii diferenţiale de ordinul al doilea. 336
§ 8. Metode cu diferenţe finite de rezolvare a ecuaţiilor cu derivate parţiale. 340
§ 9. Exerciţii. 344
BIBLIOGRAFIE. 346